Toetsing van technische vaardigheden van huisartsen : studies naar toepassingsmogelijkheden van vaardigheidstoetsing in deskundigheidsbevordering

Toetsing van technische vaardigheden van huisartsen

Studies naar toepassingsmogelijkheden van vaardigheidstoetsing in deskundigheidsbevordering

Koos Jansen

ISBN 90-5681-036-7

Vormgeving: Karin Vaessen en Maria Jansen
Ontwerp omslag: Maria Jansen
Illustraties: David Jansen
Druk: Unigraphic, Maastricht

Toetsing van technische vaardigheden van huisartsen

Studies naar toepassingsmogelijkheden van
vaardigheidstoetsing in deskundigheidsbevordering

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht op gezag van
de Rector Magnificus, Prof. Dr. A.C. Nieuwenhuijzen Kruseman
volgens het besluit van het College van Decanen
in het openbaar te verdedigen
op donderdag 4 juni 1998 om 16.00 uur

door

Koos Jansen

Promotoren:

Prof. Mr. Dr. R.P.T.M. Grol
Prof. Dr. C.P.M. van der Vleuten

Co-promotor:

Dr. J.J.E. Rethans

Beoordelingscommissie:

Prof. Dr. A.J. van der Linden (voorzitter)
Dr. P.A.J. Bouhuijs
Prof. Dr. M. de Haan (Vrije Universiteit Amsterdam)
Prof. Dr. J.C.M. Metz (Katholieke Universiteit Nijmegen)
Prof. Dr. A.C. Nieuwenhuijzen Kruseman

De studies waarover in dit proefschrift wordt gerapporteerd werden uitgevoerd in het kader van het onderzoeksprogramma van de Werkgroep Onderzoek Kwaliteit Huisartsgeneeskunde, een samenwerkingsverband tussen de Katholieke Universiteit Nijmegen en de Universiteit Maastricht.

De studies werden financieel mogelijk gemaakt door een subsidie van het Ministerie van Volksgezondheid, Welzijn en Sport. Aanvullende ondersteuning werd verkregen uit het Veldintensief Onderzoeksfonds van de Universiteit Maastricht, van de ziektekostenverzekeraars LIASS en CZ-groep. Voorts stelden diverse firma's (Skills Meducation, Pfizer-Bartlett, Limbs&Things, Laerdal, Organon, Janssen-Cilag, en Van der Kuip) materialen beschikbaar voor de cursussen.

Publicatie van dit proefschrift werd mede mogelijk gemaakt dankzij financiële bijdragen van de Landelijke Huisartsen Vereniging en de Werkgroep Onderzoek Kwaliteit (WOK) Huisartsgeneeskunde.

Inhoudsopgave

Voorwoord

Het is een bijzonder genoegen om uiteindelijk het voorwoord te schrijven. Een lange reis vindt zijn afronding. De artikelen vormen een tastbare neerslag van het onderzoek dat in de afgelopen jaren is verricht. Ze vormen echter slechts een deel van de werkelijkheid. Het voorwoord biedt me de mogelijkheid om de 'feiten' even te laten voor wat ze zijn en aandacht te vragen voor de context. Deze bood de noodzakelijke voorwaarden om tot wetenschappelijke feiten te geraken. Het biedt me ook de gelegenheid om de verschillende betrokkenen bij dit onderzoek voor het voetlicht te halen en te bedanken voor hun inzet, zonder welke dit proefschrift niet was verschenen.

Waarom ben ik er eigenlijk aan begonnen? Ik ben erin gerold. Teruggekeerd uit de tropen, enkele ervaringen rijker en ook wat illusies armer, had ik me opgegeven voor de kaderopleiding voor huisartsen. Nog wat verder studeren, leek me wel wat, want kennis is macht. De tropenopleiding bleek toch niet helemaal het goede voortraject voor deze opleiding, maar Harry Crebolder nodigde me desondanks uit om eens langs te komen: de 'tropen' schept een band. Ik vertelde nog wat rond te willen kijken voor ik me wilde gaan vestigen, misschien wat onderzoek doen. Hij zou uitkijken. Enkele dagen later belde Richard Grol of ik belangstelling had om onderzoeker te worden op een onderzoek naar technische vaardigheden van huisartsen. De beoogde onderzoeker had zich teruggetrokken. Richard dacht dat het met mijn technische vaardigheden - op grond van de tropenervaring - wel wat zou worden met het onderzoek. Enkele dagen later volgde het sollicitatiegesprek met Richard en Cees van der Vleuten. Ze wilden nog weten of ik iets waar ik aan begin ook afmaak. De rest zouden ze me wel bijbrengen. Ze hebben hun woord gehouden. Ik ook trouwens.

Maar ik ben inderdaad erg geholpen. Ik had het geluk op een rijdende trein te kunnen springen. Het onderzoek was al enkele maanden in gang gezet door Johannes Dalhuijsen en de begeleidingsgroep voor het onderzoek. Deze bestond uit Pie Bartholomeus (later opgevolgd door Albert Scherpbier), Richard Grol, Marian den Hollander (later opgevolgd door Hilde de Jong), Scheltus van Luijk, Jaap Metz, Lisa Tan, Cees van der Vleuten en Charles Verhoeff. Onder hun bezielende begeleiding kon ik snel vertrouwd raken met het onderzoek op het gebied van klinische competentie. Ik heb het altijd als een voorrecht ervaren om met deze groep, waarin een groot deel van de expertise in Nederland op het gebied van vaardigheidsonderwijs en -toetsing verenigd was, het onderzoek vorm te geven en uit te werken. Stuk voor stuk hebben ze essentiële bijdragen geleverd aan het onderzoek. En ze waren nooit te beroerd om tijdens de hectische perioden in het onderzoek bij te springen. Marian, Charles en Scheltus waren een grote hulp bij het samenstellen van de eerste kennistoets over vaardigheden. Scheltus was ook mijn steun en toeverlaat in de aanloop naar de eerste toetsafnames in Utrecht en Nijmegen in 1992, samen met Lisa. Jaap was bereid om zijn klinisch trainingscentrum op zijn kop te laten zetten voor het onderzoek, en ik mocht zelfs nog twee keer terugkomen.

Ik ben ook zeer geholpen door de onderzoeksassistentie. In de eerste periode was dat Laura Winkelman. Melanie van der Veeke volgde haar op. Haar managementcapaciteiten en energie vormden een belangrijke motor voor het tot een goed einde brengen van de laatste twee experimenten. Ook de data-invoer van deze experimenten nam zij voor een groot deel voor haar rekening. Statistische ondersteuning heb ik gekregen van Jeroen Pielage, werkzaam bij de SVUH in Utrecht. In de eerste fase van het onderzoek nam hij alles voor zijn rekening. Daarna maakte hij me vertrouwd met SPSS en met de mogelijkheden van de electronische snelweg als communicatie-middel. Zijn prompte reacties op mijn vragen en verzoeken om nadere analyses vormden een geweldige hulp.

De Stichting Verenigde Universitaire Huisartsopleidingen (SVUH) is tijdens het onderzoek een waardevolle samenwerkingspartner geweest. De werkgroep vaardigheden van het SVUH leverde de toetsstations voor het eerste onderzoek, en ook daarna heb ik nog dankbaar gebruik kunnen maken van de expertise binnen deze werkgroep, waarvan ik gedurende de onderzoeksperiode lid mocht zijn. Voor de kennistoetsen over vaardigheden mocht ik gebruik maken van het vragenbestand van het SVUH. Anneke Kramer en Just Eekhof leverden loyale ondersteuning bij de totstandkoming van de eerste kennis over vaardigheden toets, ondanks hun kritiek op het concept. Voor de uitvoering van het eerste onderzoek had ik me geen betere collega's kunnen wensen.

De vakgroep Huisartsgeneeskunde in Maastricht is een soort duiventil. Het kostte me enige tijd om gezichten, namen en functies van personen aan elkaar te koppelen. Bijzonder was dat ondanks de betrekkelijke grootschaligheid van de vakgroep er bij collega's, secretariaatsmedewerkers en management altijd de bereidheid was om mee te denken en te doen. Ik heb daar dankbaar gebruik van gemaakt. Marian Bruijstens en Anja van Bogaert hielpen om mijn eerste kennistoets in de goede lay-out te krijgen. Karin Vaessen verzorgde de lay-out van het manuscript. Manon van Haaren en Paula Rinkens waren een onmisbare steun tijdens de toetsdagen van het tweede experiment in Nijmegen. Terwijl onderzoek van onderwijs geen echt zwaartepunt in het onderzoeksprogramma vormde, was er vanuit de researchgroep voor het project een kritische belangstelling en steun die bijzonder stimulerend was. Ad de Bruyne, Saskia Mol, Floor Martens en Marijke Perquin, en later Jan-Joost Rethans en Christine Kooyman - mijn kamergenoten op de vakgroep - waren daarbij een prettig en relativerend klankbord. Ook op hen deed ik nooit een vergeefs beroep.

Zeer veel personen hebben in verschillende fasen van de uitvoering van het onderzoek belangrijke bijdragen geleverd. Paul van Aubel, Guus van de Beek, Pie Bartholomeus, Margo Beintema, Bettine Polak, Hieke Kruseman, Charles Phaff, Marjo Boumans, Bart Berden, Bert Zonneveld, Toon van Gerwen, Jacqueline Roebroeck, Saskia Mol en Jo Hendrick gaven de trainingen in de nascholingscursussen in Nijmegen en Maastricht. Bart Berden werkte ook mee aan de vergelijking van twee scoringssystemen voor reanimatie, evenals Charles Verhoeff.
Bij de praktische organisatie van nascholing en toets waren Michel Eichhorn (in Nijmegen),

en Marleen van Zandvoort en Esther Colberts (in Maastricht) betrouwbare steunpunten. Zij regelden ook de simulatiepatiënten - veelal medische studenten - die bij het onderzoek betrokken waren.
Speciale dank gaat ook uit naar de patiënten met diabetische retinopathie die bereid waren mee te werken aan de training en toetsing van fundoscopie, en de huisartsen en oogartsen die bereid waren voor dit doel onder hun patiënten te werven.
Een groot aantal huisartsen hebben bereidwillig meegewerkt als observator; vanuit de acht huisartsopleidingen bij het eerste deel van het onderzoek en daarna vanuit de vakgroep huisartsgeneeskunde in Maastricht en Nijmegen. In Nijmegen was Charles Verhoeff daarbij een waardevolle liaison. De bereidwillige medewerking van de observatoren vormde een belangrijke steun, en hun didactische kwaliteiten kwamen van pas in het geven van de feedback over de prestaties van de deelnemers.
Het meest erkentelijk ben ik echter de huisartsen-in-opleiding en praktiserende huisartsen die als deelnemer aan de diverse onderzoeken hebben meegewerkt. Zij stelden zich kwetsbaar op door zich op hun vingers te laten kijken bij het uitvoeren van vaardigheden.

Aanvankelijk was het onderzoek niet opgezet als een promotie-onderzoek. Op basis van de eerste resultaten lukte het Richard echter om er nog wat geld bij te praten. Ik had inmiddels al wel gemerkt dat onderzoek (en onderzoeker!) pas echt meetellen als het een promotie-onderzoek betreft. En ik wilde wel meetellen, ondanks waarschuwingen van deze en gene dat ik me daarmee heel wat ellende op de hals ging halen. Ze hebben gelijk gekregen, maar ik had het toch niet willen missen, al moet ik bekennen dat er ook momenten zijn geweest dat ik wenste dat ik beter naar hen geluisterd had.
Richard en Cees betoonden zich promotoren die je iedere promovendus zou toewensen. Het enthousiasme dat ze tonen voor hun vakgebieden was aanstekelijk. Ze waren altijd bereid tot advies of commentaar vanaf de opzet van het onderzoek tot en met de data-analyse. Bij het schrijven was hun ondersteuning vitaal. Mijn neiging om problemen uit te vergroten neutraliseerden ze geduldig door ze weer tot hun juiste proporties terug te brengen. Het was een genoegen om commentaar van ze te krijgen op mijn stukken.
Naast Richard en Cees heeft ook Harry Crebolder in de promotiefase waardevolle ondersteuning geboden. Hij hield in de bijeenkomsten de huisartsgeneeskundige invalshoek goed in de gaten. In zijn commentaar op mijn teksten toonde hij een goed gevoel voor helderheid in opbouw en taal. Grote indruk op me maakte hij door zijn bereidheid om ruimte te bieden voor Jan-Joost Rethans als co-promotor. Jan-Joost heeft mij een groot deel van het onderzoek intensief begeleid als praatpaal, commentator, adviseur. Ik was blij dat hij als copromotor wilde optreden.
Albert Scherpbier was mijn vijfde promotor. Hij heeft zich met een schijnbaar onuitputtelijk enthousiasme uitgesloofd gedurende dit onderzoek: als trainer, observator, simulatie-patiënt, amanuensis, etc. Zijn redactionele vaardigheden hebben me bij het schrijven van de verschillende artikelen en bij het afronden van het proefschrift enorm geholpen.
Met zo'n ondersteuning móet je wel promoveren.

JanWillem Achterbergh heeft mijn zuchten en steunen in de laatste fase van het afronden van het proefschrift meegemaakt. Ik ben gelukkig met de gastvrijheid en collegiale vriendschap die ik bij hem in de praktijk mag ervaren. In de drukte van de vaardigheidstoets en afronding van dit proefschrift was hij bereid om zijn 'vrije' dagdelen in te leveren.

Ook mijn familie heeft zich rondom dit proefschrift niet onbetuigd gelaten. Mijn schoonouders waren ten allen tijde bereid in te springen wanneer zorgtaken en andere verplichtingen weer eens een onontwarbare knoop dreigden te worden. Mijn vader las het manuscript kritisch door en kwam met verschillende suggesties die de helderheid ten goede zijn gekomen. Maria Jansen heb ik de floppy-disk in handen mogen geven om er een écht boek van te maken. David Jansen verzorgde de illustraties. Mijn belangrijkste maatje heb ik echter moeten beloven om haar níet te noemen, dus dat doe ik dan ook niet.

Ik ben blij dat het promotie-onderzoek achter de rug is. Maar voor 'een onderzoekje' blijf ik te porren. Ik heb in Tilburg een goede werkplek gevonden in een prachtig vak, met tal van uitdagingen: er is nog veel te leren, en bovenal veel te doen.

Tilburg, 21 maart 1998

Hoofdstuk 1

Inleiding

Technische vaardigheden en de huisarts

In de dagelijkse praktijk van de huisarts vormen medisch-technische verrichtingen, naast gespreksvoering, een belangrijk deel van de werkzaamheden (van Zutphen 1984). Onder technische vaardigheden worden de medisch technische vaardigheden verstaan die de huisarts gebruikt bij het uitvoeren van patiëntgebonden diagnostische of therapeutische handelingen. Zoals blijkt uit de beschrijvingen van het werk in de huisartspraktijk in het verleden (van Deen 1952; Hoogerzeil 1954; Bremer en van Westreenen 1964) en meer recente overzichten (van de Lisdonk et al. 1990; Lamberts 1991) hebben technische verrichtingen altijd een belangrijk onderdeel gevormd van het werk van de huisarts en nog steeds niet aan belang ingeboet: gemiddeld verricht de Nederlandse huisarts ruim 4.000 medisch-technische verrichtingen per 1.000 patiënten per jaar (Lamberts 1991). Ook in andere (angelsaksische) landen met een sterk ontwikkelde huisartsgeneeskunde blijken huisartsen een groot aantal verschillende diagnostische en therapeutische verrichtingen uit te voeren (Clute 1963; Spike en Veitch 1990; Heikes en Gjerde 1985).
Uit de beschrijving van het huisartsenwerk in diverse perioden wordt duidelijk dat taken en functies, en de daartoe benodigde vaardigheden, in de loop der tijden veranderen. Een goed voorbeeld hiervan vormt de verloskunde. Terwijl de huisarts in de vijftiger jaren een centrale rol vervulde bij zwangerschap en bevalling (van Deen 1952; Hoogerzeil 1954; Bremer en van Westreene 1964), is deze rol in de negentiger jaren voor het grootste deel van de huisartsen zeer beperkt (van de Lisdonk et al. 1990; Lamberts 1991).
Daarvoor in de plaats zijn andere taken en functies gekomen, onder invloed van processen als de vergrijzing van de bevolking, technologische vernieuwingen waardoor nieuwe diagnostische en therapeutische mogelijkheden (zoals peak-flow meting bij CARA-patiënten, thuiszorg-technologie) voor de eerste lijn beschikbaar komen, en beleidsverschuivingen bij overheid en zorgverzekeraars (onder meer transmurale zorg ontwikkeling), zoals recentelijk nog eens aangegeven in de rapportage van de Paritaire Werkgroep Huisartsenzorg (Anoniem 1995) en de discussienota van de LHV (Anoniem 1995). Door deze ontwikkelingen worden huisartsen steeds weer geplaatst voor nieuwe uitdagingen, waarvoor specifieke technische vaardigheden vereist zijn.
Terwijl het belang van beheersing van technische vaardigheden erkend wordt, hebben diverse onderzoeken aangegeven dat in het onderwijs niet alle relevante vaardigheden voldoende aan bod komen (Heikes en Gjerde 1985; Tan 1989; Spike en Veitch 1991). Daardoor zullen huisartsen na het voltooien van hun opleiding niet in alle opzichten goed voorbereid zijn op het adequaat uitvoeren van de diverse technische vaardigheden. Bovendien kan de beheersing van deze vaardigheden weer verloren gaan door onvoldoende uitvoering in de praktijk (Patrick 1992). Deze factoren leveren allen een bijdrage aan deficiënties in de vaardigheidsbeheersing

van huisartsen. Onderzoek naar de vaardigheidsbeheersing van huisartsen levert aanwijzingen op voor het bestaan van dergelijke deficiënties, zoals bij het onderzoek van de mammae (Campbell et al. 1991), reanimatie (Berden 1993), fundoscopie (Reenders et al. 1992), en otoscopie (Fisher en Pfleiderer 1992). Voor de kwaliteit van de gezondheidszorg vormt goede vaardigheidsbeheersing van de huisarts een belangrijke voorwaarde, omdat diagnostiek en behandeling van patiënten daarmee verbeterd kunnen worden, en onnodige verwijzingen kunnen worden voorkomen (Roland et al. 1991; Dinant et al. 1995). Dit sluit aan op het beleid gericht op de versterking van de functie van huisarts als poortwachter (Anoniem 1995).

Professionalisering van de huisartsgeneeskunde

De huisartsgeneeskunde is als afzonderlijke discipline relatief jong, en heeft na de tweede wereldoorlog een stormachtige ontwikkeling doorgemaakt, tot uitdrukking komend in de professionalisering van de beroepsgroep (de Jonge 1991). Een belangrijke stap in die professionalisering vormde de formulering van het Basistakenpakket van de huisarts (Springer 1983). Dit Takenpakket bevat een beschrijving van functies en taken waar de Nederlandse huisartsen op aanspreekbaar zijn. Er wordt ruime aandacht besteed aan de diverse medisch-technische vaardigheden die als relevant worden beschouwd voor het uitoefenen van het huisartsenvak. Genoemd takenpakket diende mede als uitgangspunt voor de huisarts-opleiding nieuwe stijl (Grol 1986).
Een tweede belangrijke stap werd eind tachtiger jaren gezet met de NHG-nota 'Naar criteria voor kwaliteit' (Anoniem 1987) en de LHV-discussienota over 'De positie van de huisarts in de toekomst' (Anoniem 1987) waarin positie, functie en taken nader worden omschreven met het oog op de kwaliteit van zorg. De behoefte bij de beroepsgroep om de positie van de huisarts in de gezondheidszorg nader te bepalen, ontstond mede onder invloed van maatschappelijke druk vanuit patiënten-organisaties en overheid om grotere helderheid te verschaffen ten aanzien van de eisen die aan huisartsen gesteld kunnen worden (van den Boogaard 1988; Ministerie van WVC 1991; Gezondheidsraad 1991). Daarmee werd de basis gelegd voor een kwaliteitssysteem voor de huisartsgeneeskunde (Grol 1991).
Dit kwaliteitssysteem bestaat uit een aantal onderdelen: richtlijnen voor effectieve en doelmatige zorg, toetsingsmethoden om vast te stellen in hoeverre de feitelijke zorg in overeenstemming is met de richtlijnen, en methoden voor kwaliteitsverbetering om de feitelijke zorg waar nodig te verbeteren.

De NHG-standaarden (Rutten en Thomas 1993; Thomas et al. 1996), die een uitwerking vormen van het Basistakenpakket, zijn belangrijke bouwstenen voor goede huisarts-geneeskundige zorg. Deze standaarden zijn te beschouwen als richtlijnen en zijn dan ook van belang als referentiepunt voor toetsing (Dalhuijsen et al. 1993) en deskundigheidsbevordering. In de recente discussienota van de LHV over het werk en de positie van de huisarts in de komende jaren wordt het belang van toetsing en deskundigheidsbevordering als methoden van

kwaliteitsverbetering nog eens onderstreept (Anoniem 1995).
In 1989 werd de Werkgroep Onderzoek Kwaliteit (WOK) opgericht om de ontwikkeling van een kwaliteitssysteem in de huisartsgeneeskunde met wetenschappelijk onderzoek te ondersteunen en te onderbouwen. Een van de aandachtsgebieden binnen dit onderzoek vormt de ontwikkeling en evaluatie van toetsingsmethoden, hetgeen is uitgewerkt in projecten op gebieden als kennis, technische vaardigheden, consultvoering en praktijkvoering. Inmiddels zijn toetsingsmethoden gevalideerd op het terrein van consultvoering (Van Thiel et al. 1991), kennis (Pollemans 1994; Van Leeuwen 1995) en praktijkvoering (Van den Hombergh et al. 1995). Daarnaast is toetsings-instrumentarium ontwikkeld voor diverse NHG-standaarden (Dalhuijsen et al. 1993).
Ook voor toetsing van technische vaardigheden werd de ontwikkeling van bruikbare toetsmethoden van belang geacht (Anoniem 1990; Grol 1991).

Vraagstellingen onderzoek

Terwijl het belang van goede technische vaardigheidsbeheersing voor de kwaliteit van het huisartsgeneeskundig handelen duidelijk is, is het inzicht in de feitelijke vaardigheidsbeheersing van praktizerende huisartsen tot op heden beperkt. Het project 'toetsing van technische vaardigheden van huisartsen' beoogde het ontwikkelen en evalueren van methoden en instrumenten voor toetsing van technische vaardigheden van huisartsen ten behoeve van deskundigheids- en kwaliteitsbevordering.

Het domein van medisch-technische vaardigheden

Voor het onderzoek was het van belang om duidelijkheid te verkrijgen welke vaardigheden beschouwd dienen te worden als huisartsgeneeskundige vaardigheden. Er bestond bij de aanvang van het onderzoek geen geactualiseerde lijst van medisch-technische vaardigheden die de huisarts zou moeten kunnen beheersen. Het Basistakenpakket gaf weliswaar een vrij gedetailleerde opsomming van vaardigheden die een huisarts diende te beheersen, maar deze lijst was, zo werd ook in het rapport gesteld, zeker niet limitatief bedoeld (Springer 1983). De meest recente lijst dateerde uit 1974 (Anoniem 1974). Deze lijst van ziekten en medisch-technische vaardigheden was wel in 1989 nog gevalideerd door huisarts-experts in het onderzoek van Tan naar tekorten in de opleiding van huisartsen (Tan 1989). Daarnaast was in Zuid-Limburg ten behoeve van het postacademisch onderwijs voor huisartsen een lijst van technische vaardigheden gehanteerd, gebaseerd op het curriculum van de Universiteit Maastricht, om regionaal de behoefte aan nascholing onder huisartsen te peilen (Bouhuijs 1981; Beusmans et al. 1985). De beschikbaarheid van een geactualiseerde lijst van technische vaardigheden was wenselijk als basis voor de keuze van onderwerpen die in toetsing en deskundigheidsbevordering prioriteit zouden moeten krijgen.
Dit leidde tot de volgende vraagstellingen:

Welke medisch-technische vaardigheden vallen binnen het domein van de Nederlandse huisartsgeneeskunde? Welke vaardigheden dienen daarbij prioriteit te krijgen in het kader van deskundigheidsbevordering en toetsing?

In hoofdstuk 2 worden deze vraagstellingen beantwoord.

Methoden van vaardigheidstoetsing

Een tweede terrein waarop onduidelijkheid bestond was de geschiktheid van verschillende methoden van toetsing voor het meten van competentie van praktizerende huisartsen ten aanzien van technische vaardigheden. In de eerder genoemde onderzoeken naar beheersing van technische vaardigheden door huisartsen werd de kwaliteit van de beheersing voor een bepaalde vaardigheid beoordeeld door het uitvoeren van de handeling te observeren. Een dergelijke keuze ligt voor de hand, omdat hiermee op directe wijze vastgesteld wordt in hoeverre een vaardigheid beheerst wordt (Metz 1984; Wakefield 1985; Patrick 1992; Miller 1990). Op het gebied van toetsing van medische competentie in het medisch onderwijs waren er ontwikkelingen waarbij het observeren van handelingen ook hernieuwde aandacht kreeg (Fabb en Marshall 1983; Neufeld en Norman 1985). Deze ontwikkelingen hadden vooral plaatsgevonden binnen de medische faculteiten en in mindere mate binnen de vervolgopleidingen, en de vraag was welke waarde deze toetsvormen bij praktizerende artsen hadden.
Duidelijk was dat observeren van een groot aantal verschillende vaardigheden kost veel toetstijd kost. De vraag was dan ook of er geen alternatieve toetsvormen zijn die een hoge voorspellende waarde hebben voor vaardigheidsbeheersing (Neufeld en Norman 1985) en eenvoudiger toe te passen zijn dan observatie. Dit leidde tot de volgende vraagstelling:

Welke methoden en instrumenten zijn geschikt voor toetsing van technische vaardigheden van huisartsen en wat zijn hun meettechnische eigenschappen?

In hoofdstuk 3 wordt de relevante literatuur over dit onderwerp besproken, en de keuze van methoden in dit proefschrift verantwoord. Er waren drie methoden die voor nadere evaluatie in aanmerking kwamen: een vaardigheidstoets (Harden 1979; Van der Vleuten en Swanson 1990), een schriftelijke kennistoets over vaardigheden (Van der Vleuten et al. 1989) en een zelfbeoordeling van beheersing van medisch-technische vaardigheden (Fuhrmann en Weissburg 1978; Dochy en Van Luijk 1987).
Vervolgens wordt in hoofdstuk 4 een eerste experiment onder huisartsen en huisartsen-in-opleiding beschreven waarin de meettechnische eigenschappen van de drie verschillende methoden worden geëvalueerd, alsmede hun onderlinge samenhang.
In hoofdstuk 5 wordt een tweede experiment onder huisartsen beschreven, waarin met name de construct-validiteit van de vaardigheidstoets werd onderzocht. Daarbij werd ook het verband tussen vaardigheidsbeheersing en kennis over vaardigheden nader onderzocht.
In hoofdstuk 6 wordt nader ingegaan op de scoring bij vaardigheidstoetsing. Voor reanimatie werden twee scoringssystemen met elkaar vergeleken: een scoringssysteem gebaseerd op een

scoringslijst die ingevuld wordt door een observator en een scoringssysteem gebaseerd op machinale registratie van de reanimatie handelingen.

De effectiviteit van nascholing en toetsing

Het verzamelen van valide en betrouwbare gegevens over kennis en vaardigheden van huisartsen vormt op zichzelf een belangrijke grond voor toetsing als basis voor mogelijke verbetering van de kwaliteit van zorg. Deze gegevens kunnen bijvoorbeeld worden gebruikt om prioriteiten te bepalen voor deskundigheidsbevordering of om de effecten van acties gericht op verbetering van competentie van huisartsen te evalueren. Vanuit de onderwijskundige literatuur wordt daarnaast ook gewezen op het potentiële leereffect van toetsing voor de individuele deelnemer (Frederiksen 1984; Newble en Jaeger 1983; Bouhuijs et al. 1987; Van der Vleuten et al. 1989). Uit de eerste twee experimenten (hoofdstuk 4 en 5) bleek, dat de resultaten van toetsing door de deelnemers als zinvolle feedback werden ervaren. De vraag was in hoeverre de meerwaarde van een dergelijke toetsing ook tot uitdrukking zou komen in een trainingseffect. In een derde experiment werd getracht enig inzicht te verkrijgen in de effectiviteit van een specifieke trainingsvorm van technische vaardigheden, waarbij toetsing een integraal onderdeel vormde van de nascholing.
Daarnaast bood dit experiment gelegenheid om de effectiviteit van feedback op zelfbeoordeling na te gaan. Op basis van de veronderstelling dat zelfbeoordeling een vaardigheid is die geleerd moet worden (Gordon 1991; Gordon 1992), was de verwachting dat herhaalde feedback zou leiden tot een sterker verband tussen zelfbeoordeling en competentie.
De volgende vraagstellingen werden geformuleerd:

Wat is de effectiviteit van nascholing en toetsing van technische vaardigheden op het handelen in de praktijk? Heeft feedback over toetsingsresultaten effect op zelfbeoordeling?

Vier verschillende technische vaardigheden werden in dit derde experiment betrokken. De resultaten van de nascholing en toetsing op zowel de competentie als het feitelijk handelen in de praktijk worden besproken in hoofdstuk 7. In hoofdstuk 8 wordt het effect van feedback op de zelfbeoordeling van technische vaardigheden besproken.

De haalbaarheid van vaardigheidstoetsing

Uiteindelijk staat of valt elk systeem van toetsing met de praktische haalbaarheid in de praktijk. De primaire invalshoek die in het onderzoek was gekozen vormde de educatieve toepassing van toetsing. Daarbij staat het mogelijk leereffect van toetsing voor de deelnemers voorop. Gezien de beperkte ervaringen in de Nederlandse huisartsgeneeskunde met toetsing (Grol en Mesker 1986; Rutten en Thomas 1993), was een vraag in hoeverre huisartsen bereid waren om zichzelf te laten toetsen en hoe deelnemers de toetsing zouden ervaren. Een tweede deelvraag betrof de wijze waarop toetsing het beste zou kunnen worden aangeboden: als geïsoleerde activiteit of als onderdeel van nascholing.
Voorts was er ook geen inzicht in de kosten en organisatie van vaardigheidstoetsing in de

Nederlandse context. In de internationale literatuur bleken kosten - in samenhang met de verschillende contexten waarin toetsing werd georganiseerd - enorm te verschillen (Reznick et al. 1993; Cusimano et al. 1994; Carpenter 1995). Dit leidde tot de volgende vraagstellingen:

Hoe is de acceptatie onder huisartsen van educatieve toetsing van technische vaardigheden? Welke vorm van toetsing heeft de meeste voorkeur? Welke zijn de kosten en organisatorische randvoorwaarden voor toetsing van vaardigheden?

Om deze vragen te beantwoorden werden tijdens de drie eerder genoemde experimenten schriftelijke enquetes afgenomen bij de deelnemers. Ook werden van alle experimenten de kosten bijgehouden. De resultaten worden behandeld in hoofdstuk 9.

Het proefschrift wordt afgesloten met een bespreking van de belangrijkste bevindingen van het onderzoeksproject, gevolgd door enkele methodologische kanttekeningen, en een aantal aanbevelingen voor onderzoek en praktijk (hoofdstuk 10).

Leeswijzer

Het proefschrift is, behoudens de inleidende hoofdstukken en beschouwing, opgebouwd uit artikelen. Herhalingen zijn daardoor onvermijdelijk. Voor een snelle oriëntatie volstaat het lezen van de samenvatting. Indien men iets meer tijd heeft dan verschaft lezing van de inleiding (hoofdstuk 1) en de beschouwing (hoofdstuk 10) een goed inzicht in de vraagstellingen van het onderzoek, de belangrijkste resultaten en de aanbevelingen voor onderzoek en praktijk.

Literatuur

Anoniem. Naar criteria voor kwaliteit. Utrecht: NHG, 1987.

Anoniem. De positie van de huisarts in de toekomst. Discussienota. Utrecht: LHV, 1987.

Anoniem. Kwaliteits- en deskundigheidsbevordering. Utrecht: LHV, 1990.

Anoniem. Werkbare wetenschap. NHG-beleidsplan 1994-1998. Utrecht: NHG, 1994.

Anoniem. Lijst van meer of minder gangbare werkzaamheden in de praktijk van de huisarts. Groningen: Buro Onderwijs Ontwikkeling Geneeskunde, 1974.

Anoniem. Poortwachter in Praktijk. Utrecht: Paritaire Werkgroep Huisartsenzorg, 1995.

Anoniem. De wereld verandert en de huisarts verandert mee.Utrecht: LHV, 1995.

Berden HJJM. Basic Cardiopulmonary Resuscitation. Assessment of skills in training situations. Dissertation. Utrecht: University of Utrecht, 1993.

Beusmans GHMJ, Verwijnen GM, Vierhout WPM, Stalenhoef PA, Van Luijk S. Evaluatie en toetsing geïntegreerd.

Een meerjarig nascholingscurriculum voor huisartsen in Limburg. Med Contact 1985;40:328-30.

Bouhuijs PAJ. Onderwerpen voor regionale nascholingscursussen. Wat vindt de huisarts er zelf van? Med Contact 1991;36:599-602.

Bouhuijs P, Van der Vleuten C, Van Luijk S. The OSCE as part of a systematic skills training approach. Med Teacher 1987;9:183-91.

Bremer GJ, van Westreenen E. De werkzaamheden in de huisartspraktijk, nu en in de toekomst. Huisarts Wet 1964;7:2-17.

Campbell HS, Fletcher SW, Lin S, Pilgrim CA, Morgan TM. Improving physicians' and nurses clinical breast examination: a randomized controlled trial. Am J Prev Med 1991;7:1-8.

Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med 1995; 70: 828-33.

Clute KF. The general practitioner. Toronto: Toronto University Press, 1963.

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE. Acad Med 1994;69:571-6.

Dalhuijsen J, Zwaard A, Grol R, Mokkink H. Het handelen van huisartsen volgens de standaard otitis media acuta van het Nederlands Huisartsen Genotschap. Ned Tijdschr Geneesk 1993;137:2139-43.

Dochy FJ, Van Luijk SJ (red). Handboek Vaardigheidsonderwijs. Lisse: Swets & Zeitlinger, 1987

De Jonge MJA. Normering van het huisartsgeneeskundig handelen: straks een gewone zaak. Huisarts Wet 1991;43:124-9.

Dinant GJ, Filion-Laporte L, Op 't Root J, Crebolder H. The Dutch advanced training programme for general practitioners; an international initiative. Eur J Gen Pr 1995;1:121-3.

Fabb WE, Marshall JR (eds). The assessment of clinical competence in general family practice. Lancaster (UK): MTP press, 1983.

Fisher EW, Pfleiderer AG. Assessment of otoscopic skills of general practitioners and medical students: is there room for improvement? Br J Gen Pract 1992;42:65-7.

Frederiksen N. The real test bias: influences of testing on teaching and learning. Am Psychol 1984;39:193-202.

Fuhrmann BS, Weissburg MJ. Self-assessment. In: Morgan KM, Irby DM (eds) Evaluating clinical competence in the health professions. St Louis: CV Mosby, 1978: 139-50.

Gezondheidsraad. Medisch handelen op een tweesprong. Den Haag, 1991.

Gordon MJ. A review ot the validity and accuarcy of self-assessments in health professions training. Acad Med 1991;66:762-9.

Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med 1992;67:672-9.

Grol RPTM, Mesker PJR. Huisarts en onderlinge toetsing. Methoden, normen en protocollen. Utrecht: Bunge, 1986.

Grol RPTM. Naar een 'kwaliteitssysteem' in de huisartsgeneeskunde.(Inaugurale rede). Utrecht: NHG, 1991.

Harden R, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE) ASME Medical education booklet no. 8. Med Educ 1979;13:41-54.

Heikes LG, Gjerde CL. Office procedural skills in family medicine. J Med Educ 1985;60:444-53.

Hogerzeil HHW. Resultaten in een huisartspraktijk. (Proefschrift). Utrecht, 1954.

Lamberts H. In het huis van de huisarts. Verslag van het Transitieproject. Lelystad: Meditekst, 1991

Metz JCM. Medische competentie. Een onderzoek naar de betrouwbaarheid en de validiteit van het Gestructureerd Klinisch Examen. (Proefschrift). Nijmegen, 1984.

Ministerie van WVC. Nota Kwaliteit van Zorg. Rijswijk: SDU, 1991.

Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-7.

Neufeld V, Norman GR (eds). Assessing clinical competence. New York: Springer, 1985.

Newble DI, Jaeger K. The effect of assessments and examinations on het learning of medical students. Med Educ 1983;17:165-71.

Patrick J. Training: Research and Practice. London: Academic Press, 1992.

Pollemans M. Kennistoetsing bij huisartsen. (Proefschrift). Maastricht, 1994.

Reenders K, De Nobel E, Van den Hoogen HJM, van Weel C. Screening for diabetic retinopathy by general practitioners. Scand J Primary Health Care 1992;10:306-9.

Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estmating the real cost of an objective structured clinical examination. Acad Med 1993;68:513-7.

Roland MO, Porter RW, Matthews JG, Redden JF, Simonds GW, Bewley B. Improving care: a study of orthopaedic outpatient referrals. Br Med J 1991;302:1124-8.

Rutten GEHM, Thomas S (red). NHG-standaarden voor de huisarts. Utrecht: Bunge, 1993.

Spike N, Veitch C. Procedural skills for general practice. Aust Fam Physician 1990;19:1545-53.

Spike N, Veitch C. Competency of medical students in general practice procedural skills. Aust Fam Physician 1991;20:586-91.

Springer MP (red). Basistakenpakket van de huisarts. Utrecht: LHV, 1983.

Tan LHC. Tekorten in de opleiding van huisartsen. (Proefschrift). Amsterdam: Universiteit van Amsterdam, 1989.

Thomas S, Geijer RMM, Van der Laan JR, Wiersma T. NHG-standaarden voor de huisarts II. Utrecht: Bunge, 1996.

Van den Boogaard CJM. Kwaliteit Huisartsgeneeskunde. Rijswijk: Geneeskundige Hoofdinspectie van de Volksgezondheid, 1988.

Van den Hombergh P, Grol R, Smits AJN, Van den Bosch WHHM, Visitatie van huisartspraktijken. Naar toetsing van de praktijkvoering. Huisarts Wet 1995;38:169-74.

Van de Lisdonk EH, van den Bosch WJHM, Huygen FJA, Lagro-Janssen ALM (red). Ziekten in de huisartspraktijk. Utrecht: Bunge,1990.

Van Deen KJ. Arbeidsanalyse in een plattelandspraktijk. (Proefschrift). Groningen, 1952.

Van der Vleuten CPM, Van Luijk SJ, Beckers HJM. A written test as an alternative to performance testing. Med Educ 1989;23:97-107.

Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teaching and Learning in Medicine 1990; 2: 58-76.

Van Leeuwen YD. Growth in knowledge of trainees in general practice. (Proefschrift). Maastricht, 1995.

Van Thiel J, Kraan HF, Van der Vleuten CPM. Reliability and feasability of measuring interviewing skills using the revised Maastricht History Taking and Advice Checklist. Med Educ 1991;25:224-9.

Van Zutphen WM. De taken van de huisarts. (Proefschrift). Maastricht: Rijksuniveristeit Limburg 1984.

Wakefield J. Direct observation. In: Neufeld V, Norman GR (eds). Assessing clinical competence. New York: Springer, 1985: 51-70.

Hoofdstuk 2

Het domein van medisch-technische vaardigheden

Inleiding

Onder medisch-technische vaardigheden worden in dit proefschrift vaardigheden verstaan die de huisarts gebruikt bij het uitvoeren van patiëntgebonden diagnostische of therapeutische handelingen. In concreto gaat het daarbij om lichamelijk onderzoek, om aanvullend onderzoek in de vorm van functietests of laboratoriumonderzoek, en om therapeutische ingrepen. Consultvoering en therapeutische interventies door middel van gespreksvoering worden niet tot de medisch-technische vaardigheden gerekend.
Er bestond ten tijde van de start van het onderzoek geen geactualiseerde lijst van medisch-technische vaardigheden die voor de huisarts relevant zijn. Het Basistakenpakket voor de huisarts (Springer 1983), dat door de beroepsgroep was geaccordeerd als een pakket waarop zij zich aanspreekbaar achtte, gaf een vrij gedetailleerde opsomming van vaardigheden die een huisarts diende te beheersen, maar deze lijst was niet volledig. Andere lijsten waren van oudere datum (Anoniem 1974), maar daarna nog wel gevalideerd (Tan 1989), of gebaseerd op het vaardigheidsonderwijs van de Universiteit Maastricht (Bouhuijs 1981; Beusmans et al. 1985). Een geactualiseerde lijst van vaardigheden, waarin ook recente ontwikkelingen in verband met thuiszorg en substitutie waren opgenomen, was nodig om de onderwerpen te kiezen die prioriteit hadden en binnen de beperkte duur van het project vertaald konden worden in toetsinstrumenten en onderwijsmateriaal.

Deze overwegingen leidden tot de volgende vraagstellingen:

Welke medisch-technische vaardigheden vallen binnen het domein van de nederlandse huisartsgeneeskunde? Welke vaardigheden dienen daarbij prioriteit te krijgen in het kader van deskundigheidsbevordering en toetsing?

Methode

Ten behoeve van het samenstellen van een lijst met medisch-technische vaardigheden die relevant zijn voor nederlandse huisartsen vond eerst een inventarisatie plaats waarbij gebruik werd gemaakt van een aantal bronnen, die hieronder kort omschreven worden. Voorts werd een inventarisatie gemaakt van vaardigheden die relevant zijn voor taken die strikt genomen niet onder het LHV-Basistakenpakket vallen, maar wel in de eerste lijn op verantwoorde wijze uitgevoerd kunnen worden. Daarbij is nadrukkelijk een aantal recente ontwikkelingen in de nederlandse gezondheidszorg (substitutie en thuiszorg) in de beschouwing meegenomen.

1. *LHV Basistakenpakket* (Springer 1983). Dit document is samengesteld in 1983 op basis van een inventarisatie van taken van de huisarts (Rapport van de commissie Takenpakket van de LHV, 1977) en commentaren op dit rapport), het rapport 'Kenmerken van de huisarts II' (Van Es et al. 1983), de Functieomschrijving van de huisarts (Anoniem 1981), de rapporten 'methodisch werken' (Anoniem 1972) en 'samenwerking, een inventarisatie' (Anoniem 1982), en het boek 'Huisarts en somatische fixatie' (Grol 1983). Het Basistakenpakket geeft een vrij globale, zeker geen limitatieve, beschrijving van de taken van de huisarts. Het is in 1983 met algemene stemmen aangenomen door de LHV-ledenvergadering als normerend voor de beroepsuitoefening, en als een 'uitstekende richtlijn (..) voor toetsing, nascholing en wetenschappelijk onderzoek'. Naast de 'basistaken' worden in de appendix ook een aantal facultatieve taken omschreven.
2. *BOOG-lijst* (Anoniem 1974). Deze lijst is in 1974 opgesteld door het Buro Onderwijs Ontwikkeling Geneeskunde in Groningen in samenwerking met het universitaire huisartsinstituut in Groningen. De lijst kwam tot stand op basis van een inventarisatie van de medisch-technische handelingen in 20 huisartspraktijken in Groningen. De lijst bevat in totaal 345 items. Deze lijst is de meest gedetailleerde beschrijving van medisch-technische vaardigheden in de huisartsgeneeskunde die momenteel beschikbaar is. De lijst is onder meer gebruikt in het onderzoek van Tan naar tekorten in de opleiding van huisartsen (Tan 1989).
3. *NHG-standaarden* (Rutten en Thomas 1993; Thomas et al. 1996). De standaarden die sinds 1989 met enige regelmaat door het NHG worden uitgebracht geven richtlijnen voor aanpak van veel voorkomende of anderszins belangwekkende problemen in de huisartsgeneeskunde. De standaarden kunnen beschouwd worden als een beschrijving van de "state of the art" van de nederlandse huisartsgeneeskunde. Daarbij richt men zich ook naar het LHV-Basistakenpakket. Momenteel zijn er ruim 60 standaarden gepubliceerd, en een aantal in voorbereiding. Technische vaardigheden vormen geen specifiek aandachtspunt binnen de standaarden, die veel meer op het beleid gericht zijn. Uit de beschrijving van het beleid in de NHG-standaarden valt echter wel op te maken welke vaardigheden noodzakelijk zijn om dit beleid goed te kunnen uitvoeren. In een aantal standaarden en bijbehorende deskundigheidsbevorderingspakketten wordt ook expliciet aandacht gegeven aan technische vaardigheden, zoals bij ulcus cruris (aanleggen compressie gradiënt verband), perifeer arterieel vaatlijden (doppler onderzoek), fluorklachten (laboratoriumonderzoek van fluor) en schouderklachten (injectietechnieken).
4. *LHV-project Specifieke Verrichtingen* (Van Heijningen et al. 1991). Dit project heeft betrekking op een twintigtal verrichtingen die gedeeltelijk vallen binnen het Basistakenpakket en gedeeltelijk bijdragen aan substitutie van specialistische zorg. Het betrof een stimuleringsproject in samenwerking met de Vereniging Nederlandse Ziekenfondsen. In het kader van het project zijn een twintigtal protocollen opgesteld ten behoeve van verschillende verrichtingen, die in de vorm van een uitgave aan de beroepsgroep beschikbaar zijn gesteld.
5. *Thuiszorg technologie* (Beijaert et al. 1993; Klein Poelhuis et al. 1987; Smeets et al. 1993). Gezien de ontwikkelingen in de bevolking en veranderingen in de gezondheidszorg is de aandacht gegroeid voor mogelijkheden van (intensieve) thuiszorg door de huisarts. Dit vraagt om specifieke kennis en vaardigheden, zoals omgaan met infuustechnieken, sondevoeding en

zuurstof-behandeling.

De lijst die resulteerde uit de inventarisatie van bovengenoemde bronnen werd bewerkt tot beschrijvingen van technische vaardigheden die een zinvolle eenheid vormen in het kader van toetsing en deskundigheidsbevordering. Deze werden geordend conform de International Classification of Primary Care (ICPC) hoofdstukken (Lamberts en Wood 1987).
Om tot een prioriteitstelling te komen voor toetsing en deskundigheidsbevordering werd door de onderzoeker uit deze lijst een eerste selectie gemaakt aan de hand van verschillende criteria. De criteria die werden gehanteerd waren incidentie en/of prevalentie van de aandoeningen waarvoor toepassing van de vaardigheid was geindiceerd, de diagnostische en/of therapeutische waarde, en de moeilijkheidsgraad van de vaardigheid. Ook werd rekening gehouden met een zekere verdeling over het totale domein, en met de prioriteiten van de beroepsgroep en beleidsmakers (overheid/zorgverzekeraars). Tot slot vormde de toetsbaarheid van de vaardigheid een reden om deze al dan niet te selecteren.
De selectie werd daarna aan twintig willekeurig geselecteerde coördinatoren van Werkgroepen Deskundigheidsbevordering Huisartsen voorgelegd, met de vraag een hiërarchische ordening aan te brengen naar behoefte aan nascholing binnen de beroepsgroep. Door middel van een schaal lopend van 0 (onbelangrijk) tot 3 (zeer belangrijk) werd de huisartsen gevraagd voor elke vaardigheid een prioriteit aan te geven. De scores van de twintig huisartsen werden opgeteld tot een somscore.

Resultaten

De inventarisatie van vaardigheden op basis van bestaande documenten leidde in eerste instantie tot een lijst met 463 technische vaardigheden die relevant geacht mogen worden voor de huisartsgeneeskundige beroepsuitoefening. Genoemde lijst was echter niet zonder meer bruikbaar als uitgangspunt voor deskundigheidsbevordering en toetsing. Sommige vaardigheden waren zeer gedetailleerd beschreven, opgesplitst in allerlei deelvaardigheden, terwijl andere vaardigheden juist op een nogal globale wijze waren omschreven. De lijst werd daarom vervolgens door de onderzoeker bewerkt tot beschrijvingen van vaardigheden die meer een zinvolle eenheid vormen in het kader van deskundigheidsbevordering en/of toetsing. Zo werden de deelvaardigheden 'inspectie schouder', 'actief en passief onderzoek schouder', 'bewegingsonderzoek schouder', 'weerstandstests schouderbewegingen', 'palpatie schouder' vervangen door de vaardigheid 'onderzoek schouder'. Anderzijds werd de vaardigheidsbeschrijving waarbij de arm als geheel was gekozen, opgedeeld in 'onderzoek elleboog' en 'onderzoek pols/hand', omdat dit wat betreft onderzoek beter aansluit op de klachten waarmee de huisarts wordt geconfronteerd. Dit resulteerde in een lijst van 263 medisch-technische vaardigheden van huisartsen (zie tabel 1), die werden geordend volgens de hoofdstukken van de International Classification of Primary Care (Lamberts en Wood 1987).

Tabel 1. Medisch-technische vaardigheden van huisartsen geordend volgens ICPC-hoofdstukken

A. Algemeen en niet gespecificeerd
onderzoek pasgeborene
constateren van de dood
in stabiele zijligging leggen
meten lengte en gewicht
meten temperatuur
injectie i.m./i.v.
infuusbehandeling
BSE

B. Bloed en bloedvormende organen
palpatie lymfeklieren
onderzoek van de milt
test van Rumpell-Leede
Hb
maken bloeduitstrijk
leucocyten telling

D. Tractus digestivus
onderzoek van het abdomen
onderzoek van de lever
onderzoek voor liesbreuk
onderzoek andere abd. hernias
onderzoek van de anus
rectaal toucher
proctoscopie
onderzoek van het gebit
onderzoek van tong en mondbodem
onderzoek van de speekselklieren
onderzoek overig mond/keel
keelwat afnemen
afname faeceskweek
faeces macroscopisch
faeces occult bloed
faeces wormeieren
perianale plakbandproef wormeieren
incisie abces mond
drainage dentogeen abces
hechten wond mond/lip
inbrengen maagsonde
maaghevelen
maagspoelen
ascitespunctie
stoma-verzorging
verwisselen a.p.-zakje
reponeren hernia
klisma toedienen

F. Oog
bepalen visus volwassene
bepalen visus kind
refractioneren
test kleurenzien
onderzoek gezichtsvelden
onderzoek van de oogstand
onderzoek oogbewegingen
afdekproeven
onderzoek convergentie-reactie
test binoculair zien
onderzoek uitwendige oog en adnexen
onderzoek cornea
onderzoek voorste oogsegment
onderzoek traansysteem
onderzoek traanproductie (Shirmer)
funduscopie
palpatie oogboldruk
tonometrie
aanbrengen oogzalf
toedienen oogdruppels
oogspoelen
verwijderen corpus al. oog
verwijderen roestring cornea
verwijderen chalazion
aanleggen oogverband

H. Oor
onderz. oorschelp en mastoid
otoscopie
gehoortest mbv spraak
horloge afstand gehoortest
onderzoek gehoor kind(Ewing)
stemvork onderzoek
audiometrie
onderzoek doorgankelijkheid gehoorgang
afnemen kweek gehoorgang
schoonmaken uitw gehoorgang
verwijderen corpus alienum uitwendige gehoorgang
toedienen oordruppels
inbrengen tampon in uitwendige gehoorgang
paracentese
evenwichtsonderzoek

K. Tractus circulatorius
onderzoek art. circul. benen (onderzoek bij TIA)
onderzoek hart
beoordeling polskwaliteiten
bloeddruk meten
onderzoek ven.circul. benen
onderzoek haemorrhoiden
EHBO arteriele bloedingen
maken en beoordelen ECG
onderzoek bij costo-clavicular compressie syndroom
sclerosering varices
incisie getrombos. haemorrh.
barronligatuur inw. haemorrh.

L. Bewegingsapparaat
onderzoek schouder
onderzoek nek
onderzoek elleboog
onderzoek pols/hand
onderzoek heupluxatie kind
onderzoek knie
onderzoek rug
onderzoek enkel
onderzoek ribben
onderzoek sleutelbeen
onderzoek voeten
onderzoek benen algemeen
therapie ganglion
punctie hydrops knie
aanlegggen mitella
aanleggen collar & cuff
aanleggen schouderverband
aanleggen zwachtel arm
aanleggen armspalk
aanleggen beenspalk
enkel tapen
aanleggen van enkelzwachtel
aanleggen gipsachterspalk
aanleggen knieverband
aanmeten hakverhoging
repositie schouderluxatie
repositie luxatie radius-kop kind
repositie vingerluxatie
repositie patella-luxatie
therapie bursitis olecrani en praepatellaris
injectie-therapie schouder
injectie-therapie elleboog
injectie-therapie pols/hand
EHBO traumatologie overig

Vervolg Tabel 1. Medisch-technische vaardigheden van huisartsen

N. Zenuwstelsel
beoordelen bewustzijn
beoordeling ontwikkeling zuigeling(Wiechen-schema)
(controle na CVA)
(controle Z. van Parkinson)
beoordelen coordinatie
beoordelen corticale functies
beoordelen motoriek
onderzoek mening. prikkeling
onderzoek spieren
onderzoek wortelprikkeling
onderzoek hersenzenuwen
orienterend reflex-onderzoek
neurologische onderzoek zuigeling en kind
neurol. onderzoek bij hoofdpijn
sensibiliteitsonderzoek
onderzoek van stamreflexen
epidurale pijnbestrijding
neurologisch onderzoek bij pijn in nek/schouder/arm
orienterend neurol. onderzoek

R. Tractus respiratorius
onderzoek longen
onderzoek neus
uitwendig onderzoek larynx
functie-onderzoek longen
onderzoek sinussen
onderzoek sputum
afnemen neuswat
toedienen neusdruppels
pleurapunctie
corpus alienum uit neus verwijderen
stelpen neusbloeding
tracheotomie
zuurstoftoediening
onderzoek tonsillen
afname keelwat
verwijderen corpus alienum uit bovenste luchtwegen
indirecte laryngoscopie

S. Huid en subcutis
onderzoek van de huid
afnemen kweek huid/abces
afnemen cytologie
afnemen histologie
maken KOH-preparaat
onderz huidparasieten (schurft)
allergie onderzoek
behandeling panaritium hand
behandeling paronychia
drainage operatie-wonden
ontlasten abces
ontlasten nagelhaematoom
aanstippen navelgranuloom
excisie atheroom
behandeling geinfecteerd cyste
excisie fibroom
excisie kleine benigne tumoren
excisie lipoom
excisie naevus + P.A.
stansbiopt
extractie nagel
fenolisatie nagelbed teen
verwijderen corpus alienum
wondhechten
partiele nagelextractie
decubitus behandeling
wondtoilet
wigexcisie teennagel
excochleatie
electrocauterisatie
cryotherapie
geleidings-anesthesie
infiltratie-anesthesie
verbandtechniek
voetverband
teenverband
handverband
vingerverband

T. Endocriene klieren / metabolisme / voeding
onderzoek schildklier
controle diabetes mellitus
glucose bloed
glucose/aceton urine
inbrengen voedingssonde

U. Urinewegen
onderzoek urinewegen
urine dipslide
urine kweek
urine onderzoek chemisch
urine macroscopisch
urine s.g.
urine sediment
inbrengen catheter man
inbrengen catheter vrouw
suprapubische blasspunctie
inbrengen/verwisselen suprapubische catheter

W. Zwangerschap / bevalling/ anticonceptie
fertiliteitsonderzoek vrouw
Sims-Huhner test
vroegdiagnostiek zwangerschap
inwendig bekken onderzoek
onderzoek zwangere uterus
begeleiding partus ontsluiting
partus uitdrijving Aav
partus uitdrijving Stuit
handgrepen schouderdystocie
episiotomie
begeleiden pasgeborene
begeleiding geboorte placenta
onderzoek ruptuur
hechten episiotomie/ruptuur
IUD plaatsen/controleren

X. Geslachtsorganen / borsten vrouw
onderzoek uitwendige genitalia
speculum-onderzoek
vaginaal toucher (bimanueel)
recto-vaginaal onderzoek
rectaal onderzoek
onderzoek borsten vrouw
gonorrhoe-diagnostiek vrouw
cervix-uitstrijk
onderzoek fluor genitalis
incisie abces bij mastitis
incisie bartholinitis
inbrengen / verwijderen / schoonmaken pessarium
cauterisatie/aanstippen erosies en condylomata
verwijderen poliepen + PA

Y. Geslachtsorg./borsten man
onderzoek uitw. genitalia man
onderzoek prostaat
gonorrhoe-diagnostiek man
scrotum-punctie
prostaat-massage
fertiliteits-onderzoek man
cauterisatie/aanstippen condylomata
sterilisatie man

Restgroep
gebruik instrumentarium
steriliasatie en desinfektie instrumentarium

Tabel 2. Prioriteitstelling onderwerpen voor toetsing en nascholing medisch technische vaardigheden

Onderwerp	Somscore (range 5-34)	Onderwerp	Somscore (range 5-34)	Onderwerp	Somscore (range 5-34)
A. Algemeen		- oog-registratie (en beoordeling)	22	* wondtoilet	15
* onderzoek pasgeborene	12	* cardiopulmonale resuscitatie	25	* hechten wond	14
- infuus-behandeling	22	- sclerosseringstherapie varices	13	- wigexcisie/tenolisatie	11
		- barronligatuur inw.haemorrhoiden	24	- behandeling ulcus cruris	26
B. Bloed en bloedvormende organen				- behandeling decubitus	26
* bepaling BSE	12	**L. Bewegingsapparaat**		- behandeling panaritium en paronychium	21
* bepaling Hb	12	* onderzoek van de rug	22	* verbandtechniek	23
		* fysische diagnostiek schouder	30		
D. Tractus digestivus		* injektie schouder	23	**T. Endocrenekliercn/metabolisme/voeding**	
* fysische diagnostiek abdomen	16	* fysische diagnostiek knie	23	* quetelet index	10
* rectaal toucher	13	* punktie hydrops knie	17	* controle diabetes mellitus	22
- proctoscopie	25	* verbandtechnieken knie	22	- inbrengen/verwisselen voedingssonde	14
- stoma-verzorging	25	* fysische diagnostiek enkel	20		
- ascitespunctie	26	* bandage/tape bij enkeldistorsie	24	**U. Urinewegen**	
		* aanbrengen van spalk	11	* urinesediment maken/beoordelen	5
F. Oog				* inbrengen catheter man/vrouw	10
* onderzoek uitwendig oog/adnexen	12	**N. Zenuwstelsel**		- inbrengen/verwisselen suprapub cath.	20
* inspektie comas (met fluorescine)	10	* beoordeling bewustzijn	16		
* onderzoek voorste oogsegment	17	* onderzoek meningeale prikkeling	12	**W. Zwangerschap/bevalling/anticonceptie**	
* onderzoek oogstand/oogbewegingen	14	* fys. diagnostiek bij HNP klachten	16	- fertiliteitsonderzoek vrouw	15
* visusonderzoek (scherpte/gezichtsv)	22	* orienterend neurologisch onderzoek	12	* onderzoek zwangere uterus	10
* fundoscopie	34	- epidurale pijnbestrijding	24	* begeleiding partus Aav	6
* tonometrie	31			* begeleiding partus stuit	8
- verwijderen roestring/corpus alienum	15	**R. Tractus respiratorius**		* plaatsen IUD	11
		* onderzoek sinussen	19		
H. Oor		* onderzoek longen	29	**X. Geslachtsorganen/borsten vrouw**	
* otoscopie	8	functie-onderzoek longen	31	* onderzoek mammae	13
* stemvorkproeven	20	* indirecte laryngoscopie	19	* vaginaal toucher/speculum onderzoek	9
- audiometrie	20	* verwijderen corpus alienum neus	15	* inbrengen/verwisselen pessarium	13
* schoonmaken gehoorgang	9	* verwijderen corpus al. bov. luchtw.	17	* cervix uitstrijk	7
* verwijderen corpus alienum	8	- toedienen O_2	34	* fluor-onderzoek	23
K. Tractus circulatorius		**S. Huid en subcutis**		**Y. Geslachtsorganen/borsten man**	
* fys.diagnostiek bij claudicatio kl.	24	* anesthesie lokaal/geleiding	18	* onderzoek prostaat	15
- doppler onderzoek bij PAV	27	* maken KOH-preparaat	21	* fertiliteitsonderzoek man	22
* tests veneuze insufficientie	22	- verwijdering huidtumor (benigne + PA)	14	- sterilisatie man	12
* fys. diagnostiek bij dec. cordis	30	* electrocauteriseren	15		
* onderzoek bij TIA's	23	* werken met vloeibare N_2	11	* Behoort tot LHV-Basistakenpakket	

De lijst omvat grotendeels vaardigheden afgeleid van het Basistakenpakket, waarvan de beroepsgroep vindt dat iedere huisarts die zou moeten beheersen (85%), en daarnaast een kleinere groep facultatieve vaardigheden (15%). De facultatieve vaardigheden zijn met name vaardigheden die verband houden met technologische vernieuwing en/of verschuiving van zorg.

Rekening houdend met de verschillende criteria werd daarna een lijst samengesteld van een tachtigtal vaardigheden, die prioriteit hadden voor toetsing en nascholing (tabel 2). De gemiddelde prioriteitsaanduiding die daaraan vervolgens werd gegeven door de twintig nascholingscoördinatoren is achter elke vaardigheid vermeld. Dit resulteerde in een top 25 aan onderwerpen voor deskundigheidsbevordering (tabel 3). Met name vaardigheden op het terrein van oogheelkundige diagnostiek, onderzoek van het bewegingsapparaat, onderzoek bij problemen van hart en vaten, en vaardigheden relevant voor thuiszorg hebben hoge prioriteit voor huisartsen die verantwoordelijk zijn voor nascholing.

Tabel 3. Top 25 vaardigheden voor deskundigheidsbevordering van huisartsen

1. Oogheelkunde	3. Thuiszorg	5. Circulatie
- fundoscopie	- infuusbehandeling	- onderzoek decompensatieklachten
- tonometrie	- pijnbestrijding	- onderzoek claudicatieklachten
- onderzoek visus	- voedingssonde	- test veneuze insufficiëntie
	- ascites-punctie	- ecg-registratie en beoordeling
2. Bewegingsapparaat	- stoma-verzorging	- cardio-pulmonale resuscitatie
- onderzoek schouder		
- injectie schouder	4. Luchtwegen	6. Dermatologie
- onderzoek knie	- fysische diagnostiek	- behandeling decubitus
- bandageren enkel	- functie-onderzoek	- behandeling ulcus cruris
- onderzoek nek/rug	- zuurstof toediening	- verbandtechniek
		7. Overige
		- proctoscopie
		- barronligatuur haemorrhoiden

Discussie

Aan de totstandkoming van de lijst met technische vaardigheden en de vaststelling van prioriteiten ligt een moeizaam proces van arbitraire keuzen ten grondslag. Allereerst geldt dat de keuze van bronnen. Weliswaar mag aangenomen worden dat met de eerder genoemde literatuurbronnen het domein grotendeels wordt gedekt, maar er kan niet gesproken worden van een uitputtende inventarisatie. Ook is er een verschil tussen lijsten die weergeven welke vaardigheden huisartsen feitelijk verrichten, op basis waarvan de de BOOG-lijst is samengesteld, en inventarisaties van vaardigheden die huisartsen zouden moeten of kunnen beheersen. De inventarisatie van technische vaardigheden van huisartsen die in dit hoofdstuk is beschreven vormt een exponent van de laatste categorie. Het zou de praktische bruikbaarheid van de lijst ten goede komen indien deze lijst nog eens zou worden vergeleken met de feitelijke praktijk.

Een volgend dilemma vormde de indeling van de vaardigheden. Geen enkel bestaand systeem bleek erg bruikbaar, zodat om pragmatische redenen gekozen werd voor aansluiting bij de International Classification of Primary Care (ICPC). Een groot gedeelte van vaardigheden kon op het nivo van de hoofdcategoriën van de ICPC-indeling eenduidig worden ondergebracht, maar er was een aantal vaardigheden dat onder twee of meer hoofdstukken ondergebracht kon worden.

De voorselectie van vaardigheden die prioriteit verdienden in het kader van nascholing en toetsing gebeurde op grond van een aantal 'criteria' die beter als aandachtspunten aangemerkt kunnen worden. Het is heel wel mogelijk dat deze voorselectie een sturend effect heeft gehad op de keuzen van de huisartsen die zich bezig houden met deskundigheidsbevordering.

Met bovengenoemde beperkingen geeft de lijst van technische vaardigheden een actueel beeld

van het brede scala aan technische verrichtingen dat door huisartsen wordt beoefend. Het maakt nog eens duidelijk dat het huisartsenvak niet alleen een praatvak, maar ook zeker een doevak is. Niet elke huisarts zal alle genoemde vaardigheden ook daadwerkelijk (willen) uitvoeren. Patiëntenpopulatie, praktijkomstandigheden, en samenwerkingsafspraken bepalen mede welke vaardigheden daadwerkelijk worden gebruikt. Het grote aandeel aan zogenaamde obligate vaardigheden maakt echter duidelijk dat huisartsen zich aanspreekbaar achten op een breed spectrum aan medisch technische handelingen.
De lijst bevat ook een aantal facultatieve vaardigheden, die vaak verband houden met vernieuwingen en verschuivingen in zorg. Verwacht mag worden dat juist op die terreinen er grote behoefte bestaat aan nascholing onder huisartsen. Uit de prioriteiten, zoals aangegeven door de WDH-coördinatoren, blijkt dat dit inderdaad het geval is. Opvallend is daarbij de hoge prioriteit die door huisartsen werd gegeven aan vaardigheden op het gebied van de oogheelkunde, het bewegingsapparaat en thuiszorg. Door verschillende huisarts-onderzoekers (Baggen 1989; Reenders et al. 1992) is gewezen op het belang voor huisartsen van deskundigheidsbevordering op het terrein van de oogheelkunde. Blijkbaar wordt die opvatting gedeeld door de WDH-coördinatoren. De aandacht voor het bewegingsapparaat is niet onbegrijpelijk gezien het grote aandeel op het spreekuur van patiënten met klachten van het bewegingsapparaat.

Wat betreft de toepassingsmogelijkheden van de lijst met technische vaardigheden voor huisartsen lijkt deze mogelijk bruikbaar als checklist voor huisartsen om voor zichzelf of met elkaar na te gaan welke vaardigheden prioriteit zouden moeten krijgen bij nascholing. Met deze vorm van behoefte peiling is eerder al in Limburg ervaring opgedaan (Bouhuijs 1981; Beusmans et al. 1985).
Omdat het huisartsgeneeskundige takenpakket en daarmee de benodigde vaardigheden aan veranderingen onderhevig zijn zou de lijst ook gebruikt kunnen worden om in een nieuwe consensus procedure het Basistakenpakket, 14 jaar na accordering, weer eens kritisch tegen het licht van de ontwikkelingen in de huisartsgeneeskunde te houden, zoals destijds overigens ook werd aanbevolen. Zowel voor beroepsopleiding als nascholing van huisartsen zou een dergelijke herijking van waarde kunnen zijn, en een stimulans kunnen betekenen om de inhoud van vaardigheidstraining opnieuw te bezien.
De keuze van prioriteiten binnen het domein van technische vaardigheden is in dat verband mede van belang, zowel in de beroepsopleiding als in nascholing. In dit onderzoek werd gekozen voor prioriteitstelling door huisartsen met een coordinerende functie in de nascholing van huisartsen, vanuit de veronderstelling dat zij goed op de hoogte zijn van de behoeften. De vraag is echter in hoeverre subjectieve prioriteiten (door huisartsen zelf ervaren nascholingsbehoefte) en objectieve prioriteiten elkaar dekken (Mast en Davis 1994). Daarom zou het wenselijk zijn om prioriteitstelling mede te laten bepalen op grond van objectieve toetsing van technische vaardigheden.

Literatuur

Anoniem. Het rapport van de commissie Takenpakket. Med Contact 1977;32:765-89.

Anoniem. Methodisch werken. Utrecht: NHG, 1978.

Anoniem. Samenwerking, een inventarisatie. Utrecht: LHV, 1982.

Anoniem. Lijst van meer of minder gangbare werkzaamheden uit de praktijk van de huisarts. Groningen: Buro Onderwijs Ontwikkeling Geneeskunde, 1974.

Anoniem.LHV-functieomschrijving van de huisarts. Med Contact 1981;36:1474-7.

Baggen JL. Oogheelkunde in de huisartspraktijk. (Proefschrift). Maastricht: Rijksuniversiteit Limburg, 1989.

Beijaert RPH, Hiemstra Y, Hoogvliet G, Lathouder HC de, Muijsenbergh METC vd, Thie J. Thuiszorg technologie. Utrecht: NHG, 1993.

Beusmans GHMJ, Verwijnen GM, Vierhout WPM, Stalenhoef PA, Van Luijk S. Evaluatie en toetsing geïntegreerd. Een meerjarig nascholingscurriculum voor huisartsen in Limburg. Med Contact 1985;40:328-30.

Bouhuijs PAJ. Onderwerpen voor regionale nascholingscursussen. Wat vindt de huisarts er zelf van? Med Contact 1981;36:599-602.

Mast T, Davis D. Concepts of competence. In: Davis DA, Fox RD (eds). The Physician as learner. Chicago: American Medical Association, 1994: 141-55.

Grol R (red). Huisarts en somatische fixatie. Utrecht: Bohn, Scheltema & Holkema, 1983.

Klein Poelhuis EH, Schadé E, Stenvers A (red). Praktische thuiszorg voor de terminale kankerpatiënt. Utrecht: Bohn, Scheltema & Holkema, 1987.

Lamberts H, Wood M. ICPC International Classification of Primary Care. Oxford: Oxford University Press, 1987.

Reenders K, De Nobel E, Van den Hoogen HJM, Van Weel C. Screening for diabetic retinopathy by general practitioners. Scand J Primary Health Care 1992;10:306-9.

Rutten GEHM, Thomas S (red). NHG-standaarden voor de huisarts. Utrecht: Bunge, 1993

Smeets PMH, Warndorff DK, Beusmans GHME. Infuusbehandeling thuis. Ervaringen met de toepassing van medische technologie. Med Contact 1993;48:905-7.

Springer MP (red). Basistakenpakket van de huisarts. Utrecht: LHV, 1983.

Tan LHC. Tekorten in de opleiding van huisartsen. (Proefschrift). Amsterdam: Universiteit van Amsterdam, 1989.

Thomas S, Geijer RMM, Van der Laan JR, Wiersma T. NHG-standaarden voor de huisarts II. Utrecht: Bunge, 1996.

Van Es JC, De Melker RA, Goosmann FCL. Kenmerken van de huisarts - II. Geheel herzien rapport onderwijsdoelstellingen van het Instituut voor Huisartsgeneeskunde van de Rijksuniversiteit Utrecht. Utrecht: Bohn, Scheltema & Holkema, 1983.

Van Heiningen JM, Hiemstra Y, Gebel RS. Protocollen specifieke verrichtingen. Utrecht: LHV, 1991.

Hoofdstuk 3

Methoden van toetsing van technische vaardigheden

Inleiding

Om iets te kunnen meten zijn meetinstrumenten nodig. Aan meetinstrumenten worden eisen gesteld van validiteit (meet de thermometer inderdaad temperatuur?), betrouwbaarheid (is het altijd 5 graden als de thermometer dat aanwijst?) en bruikbaarheid (is de temperatuur makkelijk af te lezen, en kost het apparaat niet teveel?). Voor het meten van klinische competentie kan gebruik gemaakt worden van directe toetsing, waarbij het klinisch handelen wordt geobserveerd, of indirecte toetsing. Bij indirecte toetsing wordt niet het klinisch handelen zelf gemeten, maar iets anders (bijvoorbeeld kennis) waarvan verondersteld wordt dat het een grote voorspellende waarde heeft voor klinische competentie (Miller 1993; Rethans et al. 1996). Behalve onderscheid in object van meting, kan ook nog onderscheid gemaakt worden in objectieve en subjectieve beoordelingen. Bij de laatste vorm van beoordeling is de persoon van de beoordelaar van (sterke) invloed op de uitkomst van de meting, en dat is voor meting van technische vaardigheidsbeheersing niet gewenst. Het gebruik van expliciete criteria voor de beoordeling bevordert het objectieve karakter, maar heeft als nadeel een zekere starheid en mogelijk trivialisering (Norman et al. 1991) terwijl het gebruik van impliciete criteria voor beoordeling mogelijk meer recht doet aan de complexiteit, maar het gevaar in zich heeft van een zekere willekeur (Donabedian 1986).
In dit hoofdstuk wordt een verantwoording gegeven van de methoden van toetsing die in het onderzoek zijn gebruikt om beheersing van technische vaardigheden te toetsen. Het gaat daarbij om competentie en niet om wat huisartsen met die competentie doen in de dagelijkse praktijk (Rethans et al. 1991), waarbij wel wordt verondersteld dat er een relatie is tussen kunde en toepassing. Dat die relatie overigens verre van rechtlijnig is wordt duidelijk uit de literatuur die inmiddels over dit onderwerp is verschenen (Tamblyn en Battista 1993; Davis et al. 1995, Rethans et al. 1996). De keuze van methoden kwam tot stand op basis van literatuuronderzoek en het oordeel van van deskundigen in de begeleidingsgroep van het onderzoek.* Eerst wordt de recente literatuur over medische competentie toetsing besproken, met nadruk op vaardigheidstoetsing. Daarna wordt van de toetsingsmethoden die voor dit onderzoek zijn geselecteerd de instrumentontwikkeling besproken.

---

* De deskundigen in de begeleidingsgroep waren CPM van der Vleuten en SJ van Luijk (beiden van de vakgroep Onderwijsontwikkeling en - research van de Universiteit van Maastricht), P Bartholomeus en AJJA Scherpbier (beiden van het Skillslab van de Universiteit van Maastricht), M den Hollander en H de Jong (van het Nederlands Huisartsen Genootschap), ChPM Verhoeff (vakgroep huisartsgeneeskunde van de Katholieke Universiteit van Nijmegen), JCM Metz (Klinisch Trainingscentrum van de Katholieke Universiteit van Nijmegen), RPTM Grol (Werkgroep Onderzoek Kwaliteit Huisartsgeneeskunde Katholieke Universiteit van Nijmegen/Universiteit van Maastricht).

Ontwikkeling van nieuwe vormen van toetsing

Het toetsen van medische competentie heeft in de afgelopen decennia een sterke ontwikkeling doorgemaakt (Fabb en Marshall 1983; Neufeld en Norman 1985; Van der Vleuten en Newble 1994; Van der Vleuten 1996). Een belangrijke ontwikkeling betrof nieuwe toetsvormen die de realiteit zo goed mogelijk benaderen, voortkomend uit de onvrede over de multiple choice vraagvorm die in de zestiger jaren op grote schaal werd geïntroduceerd in het onderwijs (McGuire 1987). Men had de opvatting dat de multiple choice test 'slechts' het reproduceren van feitelijke kennis toetste. Toepassing van deze kennis in de vorm van probleem oplossen of praktisch medisch handelen kon op deze wijze niet getoetst worden. Mede gevoed door nieuwe onderwijsvormen ontstonden in de 70-er jaren allerlei nieuwe instrumenten, waarin op een of andere wijze de werkelijkheid werd nagebootst. De belangrijkste exponent vormde de zogenaamde patient-management-problems (McGuire 1976). Het onderzoek naar de betrouwbaarheid dat volgde op de introductie van deze instrumenten was echter teleurstellend, want de score op een casus bleek een lage voorspellende waarde te hebben voor de score op een volgende casus, waardoor vele casus nodig zouden zijn om tot een betrouwbare score te komen (Norcini et al. 1985). Het gebruik werd dan ook afgeraden (Swanson et al. 1987). Ook modernere varianten gebaseerd op computer bleken niet wezenlijk beter (Norcini et al. 1986). Meer recent zijn echter ook een aantal nieuwe veelbelovende vormen ontwikkeld, zoals "key-feature cases" (Page et al. 1995; Bordage et al. 1995) en daarmee verband houdende casusgebonden kennistoetsing (Van Leeuwen 1994; Pollemans 1995). Ook nieuwe vormen van computersimulaties worden onderzocht (De Kock et al. 1995). Gemeenschappelijk kenmerk van deze toetsvormen is dat deze zich beperken tot essentiële elementen van een probleem, waardoor efficiënter getoetst kan worden.

Naast deze (schriftelijke) simulatie-instrumenten werden, in het streven naar meer de realiteit benaderende toetsvormen, eind zeventiger jaren observatie toetsen voor klinisch handelen ontwikkeld: de Objective Structured Clinical Examination (OSCE) (Harden en Gleeson 1979; Metz 1984) en de examenvorm gebaseerd op 'standardized-patients' (Stillman et al. 1976; Williams et al. 1987). Deze toetsen bestaan uit een aantal vaardigheden-stations, waarin kandidaten in staat worden gesteld vaardigheden te demonstreren in een gecontroleerde setting, onder directe observatie, waarbij voor de beoordeling gebruik wordt gemaakt van gestandaardiseerde scorings-formulieren (Dochy en Van Luijk 1987). Afhankelijk van de aard van de te beoordelen vaardigheden wordt daarbij gebruik gemaakt van zogenaamde 'simulatie-patiënten', ook wel 'gestandaardiseerde patiënten' genoemd (Barrows 1971; Stillman et al. 1976; Stillman et al. 1986; Ainsworth et al. 1991; Barrows 1993; Rethans et al. 1996). Deze 'patiënten' zijn geïnstrueerde leken of acteurs die op gestandaardiseerde wijze een medisch probleem presenteren. Daarnaast wordt veel gebruik gemaakt van fantomen: dat zijn lichamen of lichaamsdelen gemaakt van kunststof die zo nauwkeurig mogelijk de werkelijkheid nabootsen.
Met het gebruik van vaardighedenstations is inmiddels veel ervaring opgedaan, zowel in de

medische basisopleiding (Newble 1988; Van der Vleuten en Swanson 1990; Ainsworth et al. 1991; Harden 1992) als meer recent ook in de specialistische opleidingen (Stillman et al. 1986; Petrusa et al. 1990; Joorabchi 1991), waaronder de huisartsopleiding (Grand'Maison et al. 1992; Marshall 1993). Experimentele toepassing vindt tevens plaats in het kader van kwaliteitstoetsing van gevestigde huisartsen (Rethans et al. 1991; Norman et al. 1993; Hays et al. 1993). In Nederland wordt sinds enkele jaren vaardigheidstoetsing experimenteel toegepast in de huisartsopleiding (Tan 1988; Pollemans en Tan 1990).
De grote populariteit van deze methode, zowel bij studenten als docenten, hangt samen met de 'hoge realiteitswaarde' en 'objectiviteit' (Newble 1988; Van der Vleuten 1989; Lunenfeld et al. 1991). Onderzoek naar validiteit en betrouwbaarheid van de methode leverde over het algemeen goede resultaten op wat betreft validiteit en een redelijke tot goede betrouwbaarheid (Van der Vleuten en Swanson 1990; Colliver en Williams 1993; Vu en Barrows 1994). Met betrekking tot de betrouwbaarheid bleek met name dat de prestatie op een bepaald probleem (dat wil zeggen een bepaalde vaardigheid of een bepaalde casus) een lage voorspellende waarde heeft voor de score op een ander probleem (Norcini en Swanson 1989; Van der Vleuten en Newble 1994). Dit betekent dat slechts op basis van veel verschillende problemen een betrouwbare algemene indruk kan worden verkregen omtrent de medische competentie van een persoon. Dit gegeven blijkt overigens niet specifiek voor de vaardighedenstations, maar geldt voor alle vormen van toetsing van medische competentie (Van der Vleuten en Newble 1994). Een andere bevinding was dat er hoge correlaties werden gevonden tussen de scores op diverse vormen van toetsing, die verschillende aspecten van medische competentie beogen te meten (Van der Vleuten en Newble 1994).

De toepassing van directe observatie aan de hand van scoringslijsten om meer ervaren artsen te toetsen heeft ook kritiek ondervonden vanwege de wijze waarop klinische competentie vertaald werd in de scoringslijsten (Cox 1990; Norman et al. 1991). De scoringslijsten zouden vooral geschikt zijn om basale medisch-technische vaardigheden te toetsen, maar te rigide en/of triviaal zijn om onderscheid te kunnen maken tussen gevorderde artsen. Het beperkte valideringsonderzoek onder gevorderde artsen laat echter consistent (kleine) verschillen zien tussen groepen artsen met verschillen opleidingsniveaus (Cohen et al. 1990; Joorabchi 1991; Stillman et al. 1986; Petrusa et al. 1990), hetgeen als een ondersteuning voor de (construct) validiteit beschouwd mag worden.

Op grond van de beperkte ervaring in de huisartsopleiding in Nederland met toepassing van toetsing met behulp van vaardighedenstations is duidelijk geworden dat de methode ook in het post-academische onderwijs valide en redelijk betrouwbaar is, zij het dat aanpassingen nodig zijn in het materiaal (Tan 1988; Pollemans en Tan 1990). Specifiek op de huisartsgeneeskunde toegesneden onderwijs- en toets-materiaal wordt daarom door het Samenwerkingsverband Universitaire Huisartsopleidingen (SVUH) ontworpen en getest. Ook het NHG past in diverse deskundigheidsbevorderingspakketten scoringslijsten voor vaardigheden toe als onderwijsvorm.

Een probleem bij de toepassing van vaardighedenstations vormt de complexiteit van de organisatie en de relatief grote inspanningen die nodig zijn in termen van mensen en middelen (Newble 1988; Van der Vleuten 1989; Lunenfeld et al. 1991; Reznick et al. 1993; Cusimano et al. 1994; Carpenter 1995). Desondanks heeft in Nederland de toepassing van dergelijke toetsingsvormen geleidelijk ingang gevonden in het basiscurriculum van nagenoeg alle medische faulteiten in Nederland (Van der Vleuten et al. 1995), waarbij de toepassing het meest intensief is aan de Universiteit Maastricht. Buiten de medische faculteiten in Nederland wordt de methode slechts op beperkte experimentele schaal toegepast in de beroepsopleiding voor huisartsen (Tan 1987; Pollemans en Tan 1990). Elders in de wereld, in landen als Canada (Reznick et al. 1992; Grand'Maison et al. 1992), Australië (Hays et al. 1993) en de VS (Colliver en Williams 1993), is echter al op uitgebreidere schaal ervaring opgedaan met toepassing van de methode in zowel academisch als postacademisch onderwijs.
De complexiteit en kosten van vaardighedenstations als toetsvorm hebben geleid tot onderzoek naar goedkopere en eenvoudig toe te passen alternatieve vormen die (al dan niet gecombineerd met vaardighedenstations) gebruikt kunnen worden voor toetsing van competentie van vaardigheden. Zoals reeds opgemerkt heeft vergelijkend onderzoek tussen verschillende toetsvormen aangetoond dat er onderling hoge correlaties bestaan (Van der Vleuten en Newble 1994) en dat variantie van individuele scores tussen verschillende onderwerpen veel groter is dan tussen verschillende toetsvormen. Daarom dient getoetst te worden over een groot aantal diverse onderwerpen om een betrouwbare (reproduceerbare) individuele beoordeling te kunnen geven van vaardigheidsbeheersing. Er is op beperkte schaal onderzoek verricht naar de validiteit en betrouwbaarheid van het gebruik van een schriftelijke test om vaardigheden te meten (Van der Vleuten en van Luijk 1988). Dit onderzoek, verricht onder medische studenten aan de Universiteit Maastricht, liet in de hoogste opleidingsjaren een zeer sterke correlatie zien tussen scores op de schriftelijke toets en de vaardighedentoets. Ook in ander onderzoek werden vergelijkbare correlaties gevonden (Newble 1988; Van der Vleuten en Swanson 1990; Scherpbier 1997). De hoge correlaties die werden gevonden zouden echter ook verklaard kunnen worden uit het memoriseren van de scoringslijsten van de vaardighedentoets door studenten, waardoor bij de vaardighedenstations niet zozeer vaardigheid maar veeleer geheugenkennis werd getoetst (Van Luijk et al. 1990; Norman et al. 1991).

Een andere methode van evaluatie in het kader van vaardigheden-training en toetsing vormt het gebruik van zelf-beoordeling van de student. In de moderne onderwijstheoriën over volwassenen onderwijs vormt zelf-gestuurd leren een kernelement (Van der Vleuten 1989; Dochy en Van Luijk 1987; Knowles 1980; Schön 1987; Davis en Fox 1994). Inzicht in eigen tekortkomingen vormt in dat kader een belangrijk uitgangspunt voor keuze van leerdoelen. Zelf-beoordeling is een van de methoden die daartoe gebruikt worden. Deze opvattingen zijn ook terug te vinden in de ontwikkeling van nascholings-systemen voor de huisarts in Nederland (Beusmans 1985; Veehof 1991).
Onderzoek naar de bruikbaarheid van zelf-beoordeling levert verschillende uitkomsten op. Sommige onderzoekers vinden hoge correlaties met objectieve methoden van kennis- of

vaardigheidstoetsing (Boud en Falchikov 1989; Arnold et al. 1985), anderen echter lage correlaties (Stuart et al. 1980; Sibley et al. 1982; Kolm en Verhulst 1987; Gordon 1991).

Figuur 1 Scoringslijst Reanimatie

Scoringslijst

Toetsdatum:
Observator:
Kandidaat :

		niet gedaan	fout gedaan	goed gedaan
I. Initiële handelingen				
1.	Controleert het bewustzijn			
	- stelt met luide stem vragen aan patiënt	☐	☐	☐
	- dient sterke pijnprikkel toe	☐	☐	☐
2.	Controleert de circulatie			
	- eenzijdige palpatie van de carotis	☐	☐	☐
3.	Controleert het vrij zijn van de luchtweg	☐	☐	☐
4.	De eerste drie items worden in bovenstaande volgorde uitgevoerd	☐	☐	☐
II. Reanimatiehandelingen				
5.	Start binnen 30 seconden met reanimeren	☐	☐	☐
6.	Begint met hartmassage	☐	☐	☐
7.	De hartmassage wordt in de juiste houding uitgevoerd			
	- schouders van de arts loodrecht boven het sternum	☐	☐	☐
	- handpalmen gekruist op het sternum, twee vingers boven het os xyphoideum	☐	☐	☐
8.	De reanimatie wordt in het juiste tempo c.q. ritme uitgevoerd			
	- tempo: 80-100 thoraxcompressies per minuut	☐	☐	☐
	- ritme: afwisselend 15 thoraxcompressies en 2 beademingen	☐	☐	☐
9.	De beademing wordt technisch correct uitgevoerd			
	- het hoofd wordt in hyperextensie gebracht	☐	☐	☐
	- de mond wordt volledig over de mond van de patiënt gebracht	☐	☐	☐
	- de neus wordt dicht geknepen	☐	☐	☐
	- let op de beweging van de thorax	☐	☐	☐
10.	De thorax gaat tijdens de beademing op en neer	☐	☐	☐

Algemene indruk (0-10):

Opmerkingen:

Verschillende onderzoekers wijzen er op dat zelf-beoordeling 'geleerd' moet worden (Boud en Falchikov 1989; Calhoun et al. 1990; Gordon 1992). De bruikbaarheid lijkt echter vooralsnog vooral educatief. Met zelf-beoordeling is in Nederland ook ervaring opgedaan in de huisartsopleiding (Tan 1989; Pollemans en Tan 1990), en bij nascholing van huisartsen (Beusmans et al. 1985). Daarnaast is de methode van zelf-beoordeling ook toegepast in de medische basisopleiding om de effectiviteit van een training in vaardigheden te meten (Bleys et al. 1986).

Methoden van vaardigheidstoetsing gebruikt in het onderzoek

Op basis van de adviezen van de experts en de uitkomsten van het literatuur-onderzoek werden drie verschillende methoden in een aantal experimenten aan een nader onderzoek onderworpen.

1. een circuit van vaardighedenstations (Vaardighedentoets). Een vaardighedentoets bestaat uit een reeks van vaardighedenstations die achtereenvolgens door de kandidaten worden doorlopen. In elk station worden de handelingen van de kandidaten door observatoren beoordeeld aan de hand van scoringslijsten. De vaardighedenstations werden grotendeels ontwikkeld in de werkgroep vaardigheden van het Samenwerkingsverband Universitaire Huisartsopleidingen, bestaande uit huisartsen en toetsdeskundigen.* Een beperkt aantal stations werd buiten het verband van deze werkgroep, maar volgens een vergelijkbare procedure ontwikkeld. De stations werden waar mogelijk gebaseerd op de NHG-standaarden en/of andere huisartsgeneeskundig relevante literatuur. De stations werden in proefsituaties uitgetest voordat ze in de experimenten werden toegepast. Als voorbeeld is in figuur 1 de scoringslijst van het station 'reanimatie' weergegeven. Een lijst van onderwerpen van stations is opgenomen als bijlage 1 en 2.

2. een schriftelijke toets over vaardigheden (Kennis-over-vaardigheden toets). Deze bestaat uit een reeks van schriftelijke vragen over klinische vaardigheden. Voor de samenstelling van de kennis-over-vaardighedentoets kon voor een deel gebruik worden gemaakt van het toetsvragenbestand van de landelijke kennistoets van het Samenwerkingsverband Universitaire Huisartsopleidingen (Pollemans 1994; Van Leeuwen 1995). Dit bestand bevat een beperkt aantal vragen die kennis over vaardigheden betreffen. Daarnaast werd ook beperkt gebruik gemaakt van het vragenbestand van de Universiteit Maastricht ten behoeve van de voortgangstoets. Een groot aantal vragen werd echter ontwikkeld door de onderzoeker, ondersteund door een werkgroep met inhoudsdeskundige en toetsdeskundige expertise.** In totaal werd een bestand opgebouwd van een kleine 300 vragen. Als voorbeeld zijn in figuur 2 een aantal vragen uit de eerste KOV-toets weergegeven. Een lijst van onderwerpen voor de toetsvragen die zijn gebruikt in de diverse onderzoeken is opgenomen in bijlage 3.

3. een zelfbeoordelingsformulier t.a.v. vaardigheden (Zelfbeoordelingslijst). Deze bestaat uit een lijst van vaardigheden, waarbij de candidaten op een Likert-schaal aan kunnen geven in welke mate zij de genoemde vaardigheden denken te beheersen.
De samenstelling van deze lijsten vloeide voort uit de inhoud van de diverse experimenten. Als voorbeeld is in figuur 3 de lijst weergegeven die gebruikt werd in experiment 3.

---

* Aan de werkgroep Vaardigheden van het Samenwerkingsverband Universitaire Huisartsopleidingen werd gedurende de looptijd van het project meegewerkt door A. Kramer, J. Eekhof, L. Tan, J. Bloemen, Ch. Verhoeff, A. Chavannes, K. Habryka, M. Wieringa, B. Maiburg, S van Luijk, R. Pieters, R. Eijkelenboom en H. de Lathouder.

** Bij het samenstellen van de kennis-over-vaardigheden vragen werd ondersteuning verleend door M.den Hollander, Ch.Verhoeff, JJ. Rethans, J.Eekhof, A.Kramer, J.Bloemen en S van Luijk

Figuur 2 Voorbeelden van KOV-vragen (experiment 1)

Bij manifest strabisme kan men de scheelzienhoek schatten door middel van de zogenaamde cornea-reflex beeldjes. Valt het reflex-beeldje op de limbus (rand cornea) dan is de scheelzienhoek groter dan 30 graden. (juist)

Bij het instellen van de funduscoop dient rekening gehouden te worden met de refractieafwijking van zowel onderzoeker als patiënt (juist)

Carotismassage dient te worden uitgevoerd door dubbelzijdige druk uit te oefenen op de bulbus carotis (onjuist)

De huisarts diagnostiseert bij mevrouw Simons, 57 jaar oud, een oppervlakkige trombophlebitis op het rechter onderbeen. Hij besluit een pressure-gradiënt (compressie) verband aan te leggen. Hij legt daartoe met behulp van een elastische zwachtel van 10 cm breed een verband aan van de tenen tot net boven de knie, waarbij de hiel vrij blijft. Hij zorgt ervoor dat de druk zodanig verdeeld wordt dat deze ter plaatse van de trombophlebitis het hoogst is. Ten aanzien van de handelwijze van de huisarts gelden de volgende beweringen:

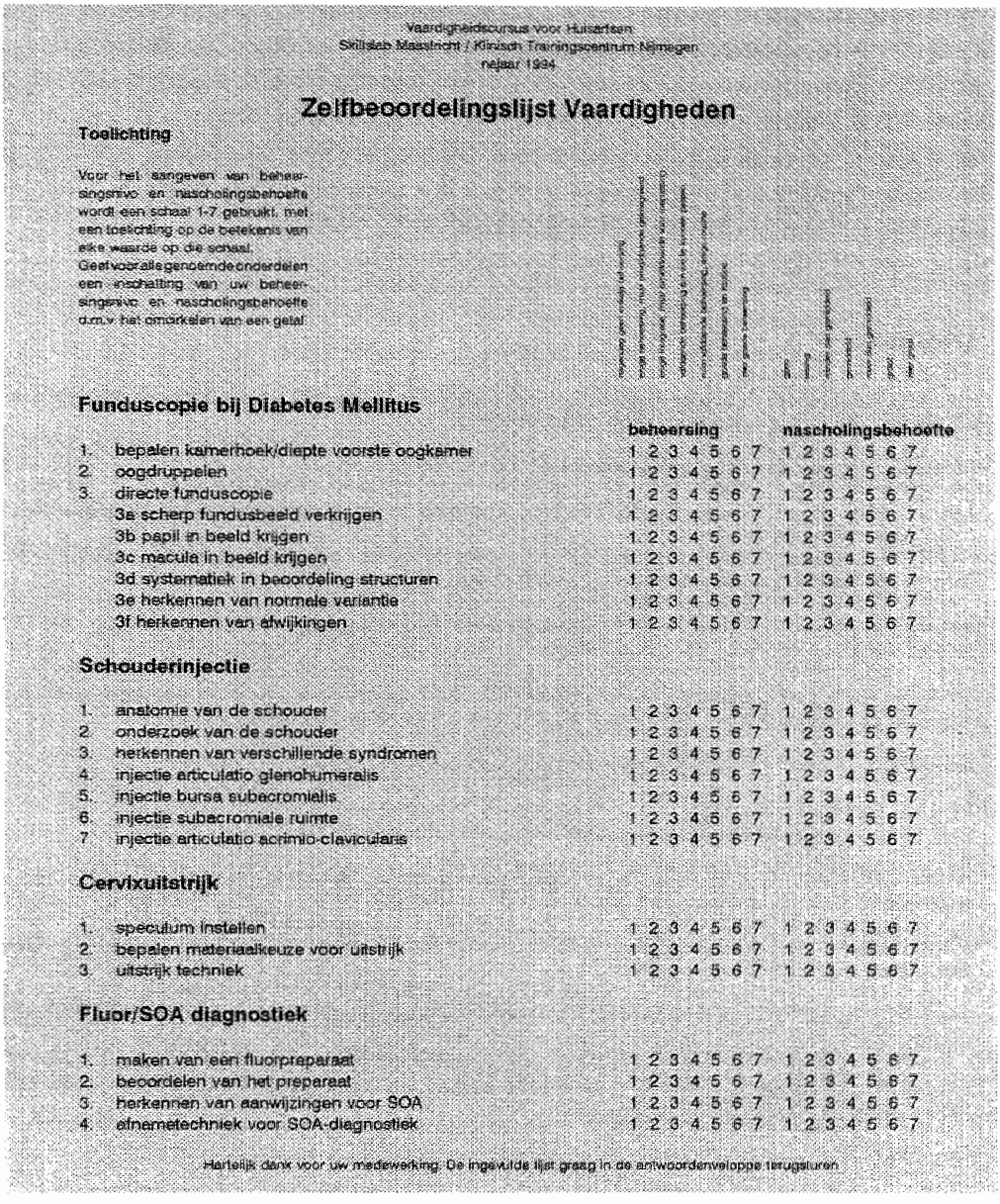
1. De gekozen zwachtel heeft de correcte breedte (juist)
2. De gekozen zwachtel is van het juiste materiaal (onjuist)
3. Het traject waarover het verband is aangelegd is correct (onjuist)
4. Het vrijlaten van de hiel is correct (onjuist)
5. De gekozen drukverdeling (ter plaatse van de trombophlebitis het hoogst) is correct (onjuist)

Discussie

Er is in de afgelopen jaren veel onderzoek gedaan naar toetsing van vaardigheden als onderdeel van de toetsing van medische competentie. Er is echter nog weinig ervaring opgedaan met toepassing van vaardigheidstoetsing bij praktizerende artsen. De aandacht richt zich bij praktizerende artsen vooral op de dagelijkse praktijk, waarbij competentie slechts een van de vele variabelen vormt in het complex van factoren die het handelen in de praktijk bepalen (Rethans et al. 1990; Grol 1992). Anderzijds wordt deskundigheidsbevordering wel in alle bestaande herregistratie systemen als een wezenlijk onderdeel beschouwd (Newble et al. 1994), waarbij toetsing van relevante aspecten van competentie in het handelen van grote betekenis worden geacht voor de verdere ontwikkeling van dergelijke systemen (Norcini 1993; Anoniem 1990).

Ten aanzien van de mogelijkheden en beperkingen van vaardigheidstoetsing onder praktizerende huisartsen bestaan er nog veel onduidelijkheden. De vaardigheidstoets, waarbij het handelen direct wordt geobserveerd, lijkt op grond van de gegevens uit de literatuur een aantrekkelijke vorm van toetsing. Er is echter twijfel over de waarde van deze vorm van toetsing bij meer ervaren artsen. Het is voorts een vrij complexe vorm van toetsing, waarvan de kosten relatief hoog zijn. Onduidelijk is ook in hoeverre deze vorm van toetsing acceptabel is voor praktizerende huisartsen.

Figuur 3 Zelfbeoordelingslijst experiment 3



Vaardigheidscursus voor Huisartsen
Skillslab Maastricht / Klinisch Trainingscentrum Nijmegen
najaar 1994

Zelfbeoordelingslijst Vaardigheden

Toelichting

Voor het aangeven van beheersingsniveau en nascholingsbehoefte wordt een schaal 1-7 gebruikt, met een toelichting op de betekenis van elke waarde op die schaal. Geef voor alle genoemde onderdelen een inschatting van uw beheersingsniveau en nascholingsbehoefte d.m.v. het omcirkelen van een getal.

	beheersing	nascholingsbehoefte
Funduscopie bij Diabetes Mellitus		
1. bepalen kamerhoek/diepte voorste oogkamer	1 2 3 4 5 6 7	1 2 3 4 5 6 7
2. oogdruppelen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3. directe funduscopie	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3a scherp fundusbeeld verkrijgen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3b papil in beeld krijgen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3c macula in beeld krijgen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3d systematiek in beoordeling structuren	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3e herkennen van normale variantie	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3f herkennen van afwijkingen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
Schouderinjectie		
1. anatomie van de schouder	1 2 3 4 5 6 7	1 2 3 4 5 6 7
2. onderzoek van de schouder	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3. herkennen van verschillende syndromen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
4. injectie articulatio glenohumeralis	1 2 3 4 5 6 7	1 2 3 4 5 6 7
5. injectie bursa subacromialis	1 2 3 4 5 6 7	1 2 3 4 5 6 7
6. injectie subacromiale ruimte	1 2 3 4 5 6 7	1 2 3 4 5 6 7
7 injectie articulatio acrimio-clavicularis	1 2 3 4 5 6 7	1 2 3 4 5 6 7
Cervixuitstrijk		
1. speculum instellen	1 2 3 4 5 6 7	1 2 3 4 5 6 7
2. bepalen materiaalkeuze voor uitstrijk	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3. uitstrijk techniek	1 2 3 4 5 6 7	1 2 3 4 5 6 7
Fluor/SOA diagnostiek		
1. maken van een fluorpreparaat	1 2 3 4 5 6 7	1 2 3 4 5 6 7
2. beoordelen van het preparaat	1 2 3 4 5 6 7	1 2 3 4 5 6 7
3. herkennen van aanwijzingen voor SOA	1 2 3 4 5 6 7	1 2 3 4 5 6 7
4. afnametechniek voor SOA-diagnostiek	1 2 3 4 5 6 7	1 2 3 4 5 6 7

Hartelijk dank voor uw medewerking. De ingevulde lijst graag in de antwoordenveloppe terugsturen.

Er lijken twee alternatieven te bestaan voor de vaardigheidstoets. De score op de kennistoets over vaardigheden blijkt een redelijke voorspellende waarde voor de score op de vaardigheidstoets te hebben, en lijkt uit oogpunt van kosten en logistiek grote voordelen te hebben voor toepassing bij grotere aantallen personen. De vraag is echter hoe sterk en consistent de relatie tussen kennis en vaardigheidsbeheersing is. Dat is met name van belang voor beoordeling van competentie voor een beperkt aantal vaardigheden, bijvoorbeeld in het kader van nascholing.

Een tweede alternatief voor de vaardigheidstoets vormt zelfbeoordeling van competentie door de huisarts. Er wordt in de literatuur over post-academisch onderwijs groot belang gehecht aan deze eigenschap, maar er is weinig onderzoek beschikbaar over de nauwkeurigheid waarmee practizerende artsen hun competentie kunnen beoordelen, en de resultaten spreken elkaar tegen. De methode is uit oogpunt van kosten en logistiek zeer aantrekkelijk.
De vragen met betrekking tot de verschillende methoden van toetsing vormden aanleiding om een aantal experimenten te organiseren waarin de drie methoden werden toegepast. Over deze experimenten wordt in de volgende hoofdstukken verslag gedaan.

Literatuur

Ainsworth MA, Rogers LP, Markus JF, Dorsey NK, Blackwell TA, Petrusa ER. Standardized patient encounters. A method for teaching and evaluation. JAMA 1991;266:1390-6.

Anoniem. Kwaliteits- en deskundigheidsbevordering. Utrecht: LHV, 1990.

Arnold L, Willoughby TL, Calkins EV. Self-evaluation in undergraduate medical education: a longitudinal perspective. J Med Educ 1985;60:21-8.

Barrows HS. Simulated patients. Springfield, Illinois: Charles C. Thomas, 1971.

Barrows HS. An overview of the use of standardized patients for teaching and evaluating clinical skills. Acad Med 1993;63:443-53.

Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP. Teaching and assessing clinical competence. Groningen Boekwerk-publications 1990.

Beusmans G. Evaluatie en toetsing geintegreerd. Med Contact 1985;11:328-30.

Bleys FC, Gerritsma JGM, Netjes I. Skills development by medical students and the influence of prior experience: a study using evaluation by student- and self-assessment. Med Educ 1986;20:234-9.

Bordage G, Brailovsky C, Carretier H, Page G. Content validation of key features on a national examination of clinical decision-making skills. Acad Med 1995;70:276-81.

Boud D, Falchikov N. Quantitative studies of student self-assessment in higher education: a critical analysis of findings. Higher Educ 1989;18:529-49.

Bouhuijs P, Van der Vleuten C, Van Luijk S. The OSCE as part of a systematic skills training approach. Med Teacher 1987;9:183-91.

Calhoun JC, Ten Haken JD, Woolliscroft JO. Medical students' development of self- and peer-assessment skills: a longitudinal study. Teaching Learning Med 1990;2:25-9.

Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med 1995;70:828-33.

Centrale Onderwijsbank. Profielbank Vaardighedenstations. Utrecht: SV-IOH, 1991.

Colliver JA, Williams RG. Technical issues: test application. Acad Med 1993;68:454-60.

Cox K. No Oscar for OSCA. Med Educ 1990;24:540-5.

Crocker L, Algina J. Introduction to classical and modern test theory. New York: Harcourt Brace Jovanovich, 1986.

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE. Acad Med 1994;69:571-6.

Davis DA, Fox RD (eds). The Physician as learner. Linking research to practice. Chicago: American Medical Association, 1994.

Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. JAMA 1995;274:700-5.

De Kock CA, Stoffers HEJH, Op 't Root JMH. De ontwikkeling van een computergestuurde casusgerichte toets voor het praktisch medisch onderwijs in de huisartsgeneeskunde. In: Pols J, et al. (red). Gezond Onderwijs 4. Houten: Bohn Stafleu van Loghum, 1995.

Dochy FJ, Van Luijk SJ (red). Handboek Vaardigheidsonderwijs. Lisse: Swets & Zeitlinger, 1987.

Dochy F, Wijnen W. Nieuwe onderwijskundige inzichten aan de basis van een moderne onderwijs-opvatting en het vaardigheidsonderwijs. in: Dochy FJ, Van Luijk SJ (red). Handboek Vaardigheidsonderwijs. Lisse: Swets & Zeitlinger, 1987:11-23.

Donabedian A. Criteria and standards for quality assessment and monitoring. QRB 1986:99-108.

Ebel RL. The practical validation of tests of ability. Educ Meas 1983;2:7-10.

Ebel RL. Must all tests be valid? Am Psychol 1961;16:640-7.

Erviti VF, Templeton B, Bunce JV, Burg FD. The relationships of pediatric resident recording behaviour across medical conditions. Med Care 1980;18:1020-1031.

Fabb WE, Marshall JR. The assessment of clinical competence in general family practice. Boston: MTP, 1983.

Frederiksen N. The real test bias: influences of testing on teaching and learning. Am Psychol 1984;39:193-202.

Gordon MJ. A review ot the validity and accuarcy of self-assessments in health professions training. Acad Med 1991;66:762-9.

Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med 1992;67:672-9.

Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective structured clinical examination for licensing of family physicians. Can Med Assoc J 1992;146:1735-40.

Grol R. Implementing guidelines in general practice care. Qual Health Care 1992;1:184-91.

Harden RM. The OSCE - a 15 year retrospective. In: Hart IR, Harden RM (eds). Current developments in assessing clinical competence. Montreal: Can-Heal Publications, 1992:41-53.

Harden R, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13:41-54.

Hays RB, Bridges-Webb C, Booth B. Quality assurance in general practice. Med Educ 1993;27:175-80.

Hessen PAW van, Verwijnen GM. De konstruktie van juist/onjuist vragen. Maastricht: PES Rijksuniversiteit Limburg, 1988.

Joorabchi B. Objective structured clinical examination in a pediatric residency program. Am J Dis Child 1991;145:757-62.

Knowles MS. The modern practice of adult education: from pedagogy to andragogy. Chicago: Follett, 1980.

Kolm P, Verhulst S. Comparing self- and supervisor evaluations. Eval Health Prof 1987;10:80-9.

Lamberts H, Wood M. The International Classification of Primary Care. Oxford Oxford University Press 1987.

Lunenfeld E, et al. Assessment of emergency medicine: a comparison of an experimental objective structured clinical examination with a practical examination. Med Educ 1991;24:38-44.

Marshall J. Assessment during postgraduate training. Acad Med 1993;68:S23-6.

McGuire C, Solomon C. Construction and use of written simulations. Chicago: The Psychological Corporation, 1976.

McGuire C. Written methods for assessing clinical competence. In: Hart I, Harden R (eds) Further developments in assessing clinical competence. Montreal: Can Heal-Publications, 1987:46-58.

Metz J. Medische Competentie. Een onderzoek naar de betrouwbaarheid en validiteit van het Gestructureerd Klinisch Examen (Proefschrift). Nijmegen: Katholieke Universiteit Nijmegen, 1984.

Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-7.

Neufeld V, Norman G (eds). Assessing Clinical Competence. New York: Springer, 1985.

Newble DI, Jaeger K. The effect of assessments and examinations on het learning of medical students. Med Educ 1983;17:165-171.

Newble DI. Eight years' experience with a structured clinical examination. Med Educ 1988;22:200-4.

Newble DI, Hoare J, Elmslie RG. The validity and reliability of a new examination of the clinical competence of medical students. Med Educ 1981;15:46-52.

Newble DI. Assessment of clinical competence: State of the art. In: Bender W, et al. Teaching and assessing clinical competence. Groningen: Boekwerk publications, 1990:23-7.

Newble D, Jolly B, Wakeford R (eds). The certification and recertification of doctors. Issues in the assessment of clinical competence. Cambridge: Cambridge University Press, 1994.

Norcini J, Meskauskas J, Langdon L, Webster G. An evaluation of a computer simulation in the assessment of physician competence. Eval Health Prof 1986;9:286-304.

Norcini J, Swanson D, Grosso L, Shea J, Webster G. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in the assessment of physician competence. Med Educ 1985;19:238-47.

Norcini J. Recertification in the medical specialties. Acad Med 1993;69:S90-4.

Norcini J, Dawson-Saunders B. Issues in recertification in North-America. In: Newble D, Jolly B, Wakeford R (eds) The certification and recertification of doctors: issues in the assessment of clinical competence. Cambridge: Cambridge University Press, 1994:28-46.

Norcini J, Swanson D. Factors influencing testing time requirements for simulation-based measurements: Do simulations ever yield reliable scores? Teach Learn Med 1989;1:85-91.

Norman GR, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Med Educ 1991;25:119-26.

Norman GR, Smith EKM, Powles ACP, Rooney PJ, Henry NL, Dodd PE. Factors underlying performance on written tests of knowledge. Med Educ 1987;21:297-304.

Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. JAMA 1993;270:1046-51.

Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. Acad Med 1995;70:194-201.

Petrusa E, Blackwell T, Ainsworth M. Reliability and validity of an objective structured clinical examination for assessing clinical performance of residents. Arch Intern Med 1990;150:573-7.

Pollemans MC, Tan LHC. Toetsing van Kwaliteit. Landelijke evaluatie van de interimbeoepsopleiding tot huisarts. Rapport SV-IOH-15. Utrecht: SV-IOH, 1990.

Pollemans MC. Kennistoetsing bij huisartsen (Proefschrift). Maastricht: Universiteit Maastricht, 1994

Rainsberry P, Grava-Gubins I, Khan S. Reliability and validity of oral examinations in family medicine. In: Hart I, Harden R (eds) Further Developments in assessing clinical competence. Montreal: Can Heal-publications, 1987:399-405.

Rethans JJ, Van Leeuwen Y, Drop M, Van der Vleuten C, Sturmans F. Competence and performance: two different

concepts in the assessment of quality of medical care. Fam Pract 1990;7:168-74.

Rethans JJ, Sturmans F, Drop R, Van der Vleuten CPM, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. Br Med J 1991;303:1377-80.

Rethans JJ, Westin S, Hays R. Methods for quality assessment in general practice. Fam Pract 1996;13:468-76.

Reznick RK, Smee S, Rothman A, Chalmers A, Swanson D, Dufresne L, et al. An Objective Structured Clinical Examination for the Licentiate; Report from the Pilot Project of the Medical Council of Canada. Acad Med 1992;67:487-94.

Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estmating the real cost of an objective structured clinical examination. Acad Med 1993;68:513-7.

Rutten GEHM, Thomas S. NHG-standaarden voor de huisarts. Utrecht: Bunge, 1993.

Scherpbier AJJA. Kwaliteit van vaardigheidsonderwijs gemeten. (Proefschrift). Maastricht: Universiteit Maastricht, 1997.

Schön DA. Educating the reflective practitioner: towards a new design for teaching and learning in the professions. San Francisco: Jossey-Bass, 1987.

Sibley JC, Sackett DL, Neufeld V, Gerrard B, Rudnick KV, Fraser W. A randomized trial of continuing medical education. New Engl J Med 1982;306:511-5.

Stillman PL, Swanson DB, Smee S, Stillman AE, Ebert TH, Emmel VE, et al.. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986;105:762-71.

Stillman PL, Sabers DL, Redfield DL. The use of paraprofessionals to teach and evaluatie interviewing skills in medical students. Pediatrics 1976;57:769-74.

Stillman P, Swanson D. Ensuring the clinical competence of medical school graduates through standardized patients. Arch Intern Med 1987;147:1049-52.

Stuart MR, Goldstein HS, Snope FC. Self-evaluation by residents in family medicine. J Fam Practice 1980;10:639-42.

Swanson D, Norcini J, Grosso L. Assessment of clinical competence: written and computer-based simulations. Assessment Eval Higher Educ 1987;12:220-46.

Tamblyn R, Battista R. Changing clinical practice: which interventions work? J Cont Educ Health Prof 1993;13:273-88.

Tan LHC, Geldorp G van, Foolen CHGM. Handleiding ter ontwikkeling van instructiestations medisch-technische vaardigheden. Utrecht, SV-IOH 1989.

Tan LHC. Studentevaluatie domein vaardigheden. Rapportage landelijke experimentele vaardighedentoets. Utrecht: SV-IOH, 1988.

Tan LHC. Tekorten in de opleiding van huisartsen. Ziektebeelden en medisch-technische vaardigheden. (Proefschrift). Amsterdam: Universiteit van Amsterdam, 1989.

Van der Vleuten CPM, Van Luijk SJ, Beckers HJM. A written test as an alternative to performance testing. Med Educ 1988;23:97-107

Van Leeuwen YD. Growth in knowledge of trainees in general practice. (Proefschrift). Maastricht: Universiteit Maastricht, 1995.

Van Luijk SJ. Al doende leert men. Enkele studies naar aspecten van betrouwbaarheid en validiteit over de toetsing van vaardigheden. (Proefschrift). Maastricht: Universiteit Maastricht, 1994.

Van Luijk SJ, Van der Vleuten CPM, Van Schelven SM. The relationship between content and psychometric characteristics in performance-based testing. Med Educ 1990;23:97-107.

Van der Vleuten CPM. Naar een rationeel systeem voor toetsing van studieprestaties in probleemgestuurd medisch onderwijs. Studies naar betrouwbaarheid en validiteit van toetsen voor praktische vaardigheden (Proefschrift).

Maastricht: Universiteit Maastricht, 1989.

Van der Vleuten C, Newble D (eds). Methods of assessment in certification. In: Newble D, Jolly B, Wakeford R (eds). The certification and recertification of doctors. Issues in the assessment of clinical competence. Cambridge: Cambridge University Press, 1994:105-125.

Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med 1990;2:58-76.

Van der Vleuten CPM, Scherpbier AJJA, Van Luijk SJ. Use of OSCEs in The Netherlands. In: Rothman AI, Cohen R (eds). The Sixth Ottawa Conference on Medical Education. Toronto: University of Toronto, 1995:320-1.

Van der Vleuten CPM. The Assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ 1996;1:41-67.

Veehof L. Curriculum voor de nascholing: steun of keurslijf. De Huisarts 1991;sep:16-8.

Watson A, Houston IB, Close GC. Evaluation of an objective structured clinical examination. Arch Dis Child 1982;57:390-8.

Williams RG, Barrows HS, Vu NV, Verhulst SJ, Colliver JA, Marcy M, et al. Direct, standardized assessment of clinical competence. Med Educ 1987;21:482-9.

Assessment of competence in technical clinical skills of general practitioners using different methods*

Summary

As technical clinical procedures constitute an important part of the work of general practitioners, assessment of competence in these skills is considered important from the perspective of quality assurance. In this study the psychometric characteristics of three different methods for assessment of competence in technical clinical skills in general practice were evaluated.
A performance-based test (8 stations), a written knowledge test of skills (125 items) and a self-assessment questionnaire (41 items) on technical clinical skills were administered to 49 general practitioners and 47 trainees in general practice.
The mean scores on the performance-based test and the written knowledge test of skills showed no substantial differences between general practitioners and trainees, whereas the general practitioners scored higher on the self-assessment questionnaire. While the correlation of the score on the knowledge test of skills with the score on the performance-based test was moderately high, the score on the self-assessment questionnaire showed a rather low correlation with the performance-based test.
Although performance-based testing is obviously the best method to assess proficiency in hands-on skills, a written test can serve as a reasonable alternative, particularly for screening and research purposes.

Introduction

Technical clinical procedures constitute an important part of the daily work of physicians (Lamberts *et al.* 1991), and proficiency in technical clinical skills is considered a relevant aspect of clinical competence (Fabb 1983). From the perspective of quality assurance of medical care it is therefore important to gather reliable data on competence in relevant technical clinical skills, as a basis for planning continuing education programs (Berg 1979). The aim of the study presented here was to identify and evaluate different methods for assessment of competence in technical clinical skills in general practice.
Direct observation of performance under standardized conditions has been identified as the first choice assessment method. This method, originally described by Harden and Gleeson (1979) as the objective structured clinical examination (OSCE), has been intensively researched, mainly in undergraduate programs and to a lesser degree in postgraduate education (Hart *et*

* Published as: Jansen JJM, Tan LHC, Van der Vleuten CPM, Van Luijk SJ, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners. Med Educ 1995; 29: 247-53.

al. 1986; Hart & Harden 1987; Bender *et al.* 1990; Hart *et al.* 1992). It has generally been considered a valuable method, because of good validity. The OSCE however has some disadvantages in terms of organizational complexity and resources needed (Anderson & Kassebaum 1993; Reznick *et al.* 1993). This threatens feasibility for widespread use in postgraduate quality assurance schemes. The use of a written test and self-assessment were therefore considered as potential alternative methods for performance-based testing.

Theoretically a relationship is assumed between knowledge and competence in skills (Patrick 1992). At graduate level the correlation between scores on performance-based tests and written tests assessing clinical competence seems variable (Van der Vleuten & Swanson 1990). Some of differences found can perhaps be explained by differences in content of the tests compared. Newble and Swanson (1988) reported a moderately high correlation (0.88) between an objective structured clinical examination (patient stations) and a short answer test in the final year examination, using the same blueprint for both tests. Van der Vleuten *et al.* (1988) also found a high correlation (0.89) between a written test and a performance-based test constructed according the same blueprint among senior medical students. These studies showed that a written test score has potential predictive value for a performance-based test score in a population of graduating students. This could however be quite different among practising physicians working in variable practice conditions and having variable continuing medical education experience.

The consideration of self-assessment as another alternative method originated from the literature on adult learning (Fuhrmann & Weissburg 1978), which views self-assessment as an important requisite for effective learning. Yet little research has been published concerning the validity of self-assessment. Results show low to moderate correlations between self-assessment and objective methods, with higher correlations between self-assessment and performance compared to self-assessment and knowledge (Gordon 1991; Gordon 1992). Most research is based on undergraduate student-populations. As self-assessment is considered a skill which has to be acquired, experienced professionals might be more accurate than undergraduate students in self-assessment of their performance of technical clinical skills (Wooliscroft *et al.* 1993).

The specific research-questions of the study presented here were:

1. Do the identified methods to assess competence in technical clinical skills discriminate between different levels of experience among general practitioners?
2. What is the reliability of the three different test methods?
3. What is the relationship of the scores of the written test and the self-assessment questionnaire with the performance-based score?

Methods

Subjects

In March 1992 a test was administered to 49 general practitioners - all involved as teachers in vocational training in general practice - and 47 trainees in general practice, recruited from two university training institutions. The general practitioners' experience in practice ranged from 5 to 25 years (mean 13 years): 11 had less than 10 years experience, 23 had 10 to 15 years experience and the remaining 15 had more than 15 years of working experience as a general practitioner. The trainees were at different stages of vocational training: 12 at 3-months (beginners), 16 at 7-months (intermediate), and 19 between 19 and 23 months (advanced).

Instruments

The test consisted of three different parts: a performance-based test, a written knowledge of skills test and a self-assessment questionnaire.

Performance-based test (PBT). Stations for the performance-based test (PBT) were developed by a national committee of six practising general practitioners and two test experts, and based on nationally accepted reference literature for general practitioners, including national guidelines for general practitioners as developed by the Dutch College of General Practitioners (Grol 1990). Checklists for scoring contained items considered as crucial for adequate performance, as agreed upon by consensus by the committee. Each item was defined in one or more subitems. An illustration is provided in Figure 1. Performance on all subitems had to be adequate to obtain a favorable marking of the item. Each correct item was given one credit point. Incorrectly or not performed items received no points.

The testing time per station varied between 10 and 20 minutes, adding to a total testing time of two hours for the eight stations. Four stations included the management of clinical problems (chest pain, urinary tract infection, impaired hearing, ankle sprain), and standardized patients were used. The remaining four stations covered the performance of isolated technical skills (ophthalmoscopy, urinary catheterisation, resuscitation, insertion of an intrauterine device). Mannequins were used in these stations. There were no written (follow-up) stations.

Knowledge test of skills (KTS). The written knowledge test of skills contained 125 items concerning different technical clinical skills relevant to general practice. The items had the form of statements requiring judgement as true or false (see Figure 1). If in doubt about the correct answer, a question mark could be used. Of the 125 items, 75 were constructed with a content corresponding to the performance-based test. The remaining 50 questions focussed on relevant technical skills not covered by the performance-based test, thus allowing comparison of prediction of the performance-based scores by the different subsets of items.

Self-assessment questionnaire. The self-assessment questionnaire (SAQ) consisted of 41 items, with 20 items corresponding to the content of the PBT. The remaining items corresponded to skills only covered by the written test. For each item the candidates were prompted to indicate the level of their proficiency in the particular skill using a 5-point Likert-scale (very poor-poor-regular-good-very good) (see Fig. 1).

Figure 1 Scoring grid, questions and self-assesment items on Resuscitation

Scoring grid Resuscitation

Testing date:
Ratercode:
Candidate:

Initial procedures	not performed	incorrect	correct
1. Checks consciousnes - tries to wake patient with loud voice - gives adequate painstimulus	☐	☐	☐
2. Checks circulation - onesided feeling for carotid-pulsations	☐	☐	☐
3. Checks if airway is free	☐	☐	☐
4. The first 3 items are performed within 30 seconds	☐	☐	☐
5. The first three items are performed in the above mentioned order	☐	☐	☐
Resuscitation			
6. Starts directly with resuscitation	☐	☐	☐
7. Cardiac massage is performed correctly - shoulders of resuscitator are above sternum patient - hands are crossed on the sternum two fingers above xyphoid - rythm: 15 compressions in 10 seconds	☐	☐	☐
8. Performs two insufflations after each 15 compressions	☐	☐	☐
9. Performs insufflations correctly - brings head patient in hyperextension - fully covers mouth patient during insufflation - doesn't allow air to escape from nose patient - watches whether chest rises during insufflation - chest rises during insufflation	☐	☐	☐

Questions on resuscitation

The GP decides to resuscitate an infant (less than 1 year old), who has no signs of spontaneous breathing nor arterial pulsations. The head of the infant is hyperextended. The correct extent of hyperextension of the head is LESS with an infant compared to an adult

The GP places his mouth over the nose and mouth of the infant. This is a correct procedure for insufflation of an infant.

During the resuscitation the GP gives thorax compression at a rate of about 90 per minute.This is a correct rate for infants.

During resuscitation of an adult the adequate rate of compressions is closer to 80 per minute than to 60 per minute.

During resuscitation with one resuscitator the recommended schedule is: 15 compressions followed by 1 insufflation.

Self-assessment questionnaire

		how often performed						proficiency					
		0-1x per year	2-6x per year	7-15x per year	16-50x per year	> 50x per year	not relevant for gp	very bad	bad	regular	good	very good	
14.	Resuscitation adult	1	2	3	4	5	6	1	2	3	4	5	
15.	Resuscitation child	1	2	3	4	5	6	1	2	3	4	5	

Procedure

The candidates were tested on four different days, and on two different sites. After the self-assessment questionnaire was completed, the candidates passed through the first part of the performance-based test, subsequently the knowledge test of skills, and finally the second part of the performance-based test. This test sequence was used for logistical reasons.

A group feedback session was held at the end. Candidates and raters were prompted to comment on the content of the test and the testing procedure was evaluated. As a result of comments by the candidates and raters 6 out of 78 items were removed from the checklists of the PBT before final analysis, and 10 items were removed from the KTS, leaving 115 items for analysis. As a consequence of the link between the items on the knowledge test and the self-assessment questionnaire, 6 items were removed from the questionnaire, having 35 items remaining on the SAQ.

The standardized patients were recruited from a group of experienced standardized patients from one of the participating universities. They were trained by a general practitioner experienced in the training of standardized patients. A total of 36 general practitioners (staff-members of departments of general practice) were involved as raters. One third of the encounters were double-rated. One week before the test the raters received a two hour training. During the training-session scoring was practised and results were compared and discussed. The interrater reliability was 0.82 for the total checklist scores (intraclass correlation, including absolute and relative differences between raters in the error term).

Analysis

For the PBT the individual testscore was calculated as the mean of the scores on the different stations. The KTS score was based on the sum of correct answers, and the score on the SAQ was calculated as the sum of scores on the Likert-5 point scale. All scores were expressed as percentages of the maximum score.

The statistical analysis performed included a one-way analysis of variance using a multiple comparisons test (Student-Neuman-Keuls) for differences between groups. Generalizability theory (Cronbach *et al.* 1972) was used to calculate the reliability coefficients for relative and absolute decisions, and interrater reliability. Correlations were calculated as Pearson Product-Moment coefficients.

Generalizability theory may be considered as an extension of classical test theory. In classical test theory the observed variance is seen as composed by two sources: true score variance and error-variance. Reliability is defined as the ratio between the true score variance and error score variance. Generalizability analysis allows for partition of the error variance into multiple sources. Depending on the perspective (relative or absolute interpretations), multiple error variances can be estimated resulting in multiple reliability coefficients. The norm-referenced reliability coefficient is appropriate when test scores are used for the rank ordering of the candidates (e.g. candidate A is better than candidate B). The domain-referenced reliability coefficient is appropriate for absolute score interpretations (e.g. candidate A masters 70% and candidate B 60%).

Table 1. Scores on the performance-based test (PBT), the knowledge test of skills (KTS) and the self-assessment questionnaire (SAQ)

	PBT Total score (8 items)			KTS total score (115 items)			KTS Subscore (72 items)			SAQ total score (35 items)			SAQ subscore (21 item)		
	mean	SD	range	mean	SD	range	mean	SD	range	mean	SD	range	mean	SD	range
trainees (n=47)	56	9	35-76	61	7	44-79	58	6	44-74	64	7	50-82	61	7	45-84
gp's (n=48)	56	9	34-75	65[a]	7	51-79	61	7	46-78	70[c]	9	53-88	69[c]	9	49-89
trainees															
beginners (n=12)	53	10	35-76	56[b]	7	44-68	56	6	44-63	63	7	50-75	60	7	45-72
intermediate (n=16)	57	8	41-73	61	5	52-71	58	5	50-68	64	5	54-75	61	7	51-74
advanced (n=19)	58	8	39-72	63	7	52-78	60	6	49-74	65	7	55-82	64	8	54-84
gp's															
< 10 yr (n=11)	56	7	44-67	65	8	54-77	61	8	49-74	68	7	60-84	65	7	45-72
10-15 yr (n=23)	55	9	34-72	64	6	51-74	60	7	46-72	71	9	53-86	69	9	51-74
15 yr < (n=14)	57	9	41-75	67	7	54-79	61	8	50-78	72	10	58-88	72	10	54-84

note: all entries are expressed as percentage scores. [a] gp's > trainees $p<0.001$ [b] beginners < intermediate = advanced $p<0.05$ [c] gp's > trainees $p<0.001$

Results

Scores

Complete data were available from all 96 candidates on the performance-based test and on the knowledge test of skills. One candidate failed to complete the self-assessment questionnaire. There were no statistically significant differences between sites and days of administration.
Table 1 shows the scores of the candidates on the performance-based-test (PTB), the written knowledge test of skills (KTS) and the self-assessment questionnaire (SAQ). Results of experienced general practitioners (GP's) were compared with trainees. Within both groups the results were broken down for differences in experience.
On the performance-based test there was no difference in mean scores between GP's and trainees. There was also no difference in score among the GP's with different years of practice experience. Within the group of trainees there was a trend of slight improvement in scores in relation to stage of vocational training.
The results on the knowledge test showed a statistically significant difference between the mean scores of GP's and trainees ($p < 0.001$). There was no difference in score between general practitioners with varying years of experience in practice. The mean scores of the trainees increased with experience-level, reaching a statistically significant difference only for the scores of the beginners-group versus the scores of the two other groups ($p < 0.05$). The subscores on the knowledge test of skills, based on the answers on the 75 questions linked with the performance-based test, however showed no statistically significant differences.
On the self-assessment questionnaire there was a significant difference between GP's and trainees for the total score as well as for the subscore ($p < 0.001$). Within the trainee-group as well as within the group of GP's there was a small increase in score associated with level of training respectively years of practice. The differences however were not statistically significant ($p > 0.05$).

Reliability

In Table 2 the results are presented of the generalizability analysis based on the personal scores. The norm-referenced reliability coefficient reflects the reliability of the rank ordering of candidates. A reliability of 0.80 is often considered as a minimum requirement if scores are used as a basis for individual decision making. The required testingtime to reach such a norm-referenced reliability was calculated for the different tests, resulting in considerable time required for the performance-based test. The domain-referenced reliability coefficient indicates how reliable the absolute scores are. It is naturally more severe since not only the differences in rank ordering but also the absolute differences in scores on the items (item- or test-difficulty) are taken into account. This explains why the required testingtime to reach a reliability coefficient of 0.80 is considerably longer compared to the norm-referenced approach. Table 2 also includes the standard error of measurement (SEM) for the different tests as an alternative reliability index. The SEM can be used to estimate a confidence interval for individual test scores (multiplying the SEM by 1.96 a 95% confidence interval is obtained, e.g. the 95%

confidence interval for the KTS-score of candidate A with a test score of 70% ranges from 61-79%). As can be seen large confidence intervals are to be taken into account for the performance-based test.

Table 2 Reliability indicators of the performance-based test (pbt), knowledge test of skills (kts) and self-assessment questionnaire (saq) for total scores and subscores

	norm-referenced reliability	testing time to reach 0.80 (hours)	domain-referenced reliability	testing time to reach 0.80 (hours)	SEM* (%)
PBT	0.43	10.0	0.35	14.5	7.7
KTS	0.68	1.7	0.64	2.1	4.5
KTS-subscore	0.43		0.37		5.8
SAQ	0.92	0.1	0.90	0.1	2.8
SAQ-subscore	0.87		0.83		3.8

* Standard error of measurement (SEM) expressed as percentage of maximum score

Correlations

The correlations between total test scores on the different assessment-methods were calculated. Calculations were repeated using the subscores of 75 items of the KTS and 19 items of the SAQ linked to the content of the PBT. The correlations were recalculated after correction for attenuation caused by the unreliability of the tests, as this tends to obscure existing relations between scores. These disattenuated correlations are indicative for the correlations which would result when the tests used would have perfect reliabilities. Results are presented in Table 3.

Table 3. Correlations between (sub)scores of the performance-based test (PBT), the knowledge test of skills (KTS) and the self-assessment questionnaire (SAQ)

	PBT	KTS	KTS sub	SAQ	SAQ sub
PBT		0.54	0.77	0.40	0.47
KTS	0.29[b]		1.00	0.37	0.40
KTS-sub	0.33[b]	0.87[c]		0.38	0.49
SAQ	0.25[a]	0.29[b]	0.24[a]		1.00
SAQ-sub	0.29[b]	0.31[b]	0.30[b]	0.96[c]	

Note: Observed correlations in lower triangle ([a] $p<0.05$, [b] $p<0.01$, [c] $p<0.001$) and disattenuated correlations in upper triangle.

The observed correlations between PBT and KTS were low, with a slightly stronger correlation of the subscores compared to total scores. The same relation can be seen between PBT and SAQ. The correlations between KTS and SAQ were within the same range. However, correcting the scores for unreliability gave moderate to high disattenuated correlations between PBT and KTS, somewhat lower correlations between the PBT and SAQ, and even lower correlations between KTS and SAQ.

Discussion

Although the results do show some small differences in mean scores between practising general practitioners and trainees, the overall results on the performance-based test and knowledge test indicate that competence in technical clinical skills (as measured by the written test or the performance-based test) shows no substantial differences. Only on the self-assessment score do trainees and general practitioners differ consistently.

The proficiency in technical clinical skills seems to show little general improvement or deterioration during vocational training and thereafter, whereas the higher self-assessment score associated with more advanced levels of training or experience most likely reflects a general self-attribution: as a result of experience general practitioners tend to feel more confident about their competence concerning technical clinical skills, without necessarily being more competent.

It has been difficult to demonstrate changes in scores on performance-based tests related to training or experience at postgraduate level, whereas these changes can easily be demonstrated on written tests (Norman *et al.* 1994). This questions the validity of the use of performance-based tests to discriminate between different degrees of expertise among general practitioners. However, as the scores on the written test also showed no substantial differences related to experience, we believe the scores on the performance-based test reflect the absence of substantial differences of competence between groups with different training-level and experience.

The results of the reliability analyses were comparable with results in the literature, taking testing time into account (Van der Vleuten & Swanson 1990). Testing time was for all but one test too short to obtain reproducible scores. The high reliability of the self-assessment questionnaire reflects the strong influence of global self-attributions (Gordon 1991).

There was a positive correlation between knowledge of skills and proficiency on these skills. The existence of this specific relation is supported by the finding of a higher correlation, linking the subscores of the test. These findings indicate that a written knowledge test of skills can predict performance on these skills to some extent, if developed according the same blueprint. This implies that a written test might be useful in situations where performance-based tests are difficult to apply, e.g. for screening purposes. The performance-based test could then be reserved for a (smaller) group identified to merit further evaluation, and thus a more

efficient use of the performance-based test is achieved.
The correlation between self-assessment and proficiency in technical skills was moderate. Other studies reported low to absent correlations between self-assessment and objective assessment methods (Gordon 1991; Stillman *et al.* 1990; Stillman *et al.* 1986). However, in contrast to the written test, the subscore of self-assessment showed only a slightly higher correlation with the performance-based test, suggesting that general practitioners have a rather general notion about their proficiency in technical clinical skills. It would be interesting to investigate whether a training program in self-assessment could improve this skill (Gordon 1992).
In conclusion, while performance-based testing is obviously the best method to assess proficiency in hands on skills, a written test can serve as a reasonable alternative in some situations, as it is relatively easy to administer and not very costly. Self-assessment, although positively correlated with performance, is a less viable alternative as it seems to reflect a general notion of competency.

References

Anderson MB, Kassebaum DG (eds) Proceedings of the AAMC's Consensus conference on the Use of Standardized Patients in the Teaching and Evaluation of Clinical Skills. Acad Med 1993;68:437-83.

Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds) Teaching and Assessing Clinical Competence. Groningen: BoekWerk Publications, 1990.

Berg AO. Does continuing medical education improve the quality of medical care? A look at the evidence. J Fam Practice 1979;8:1171-4.

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The Dependability of Behavioural Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons Inc, 1972.

Fabb WE. The assessment of clinical competence in general family practice. Hingham (USA):MTP-press, 1983.

Fuhrmann BS, Weissburg MJ. Self-assessment. In: Morgan KM, Irby DM (eds) Evaluating clinical competence in the health professions. St Louis: CV Mosby, 1978:139-150.

Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 1991;66,762-9.

Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med 1992;67:672-9.

Grol RPTM. National standard setting for quality of care in general practice: attitudes of general practitioners and response to a set of standards. Br J Gen Pr 1990;40:361-4.

Harden RM, Gleeson FA. Assessment of clinical competence using an obsjective structured clinical examination (OSCE). Med Educ 1979;13:41-54.

Hart IR, Harden RM, Des Marchais J (eds) Current Developments in Assessing Clinical Competence. Montreal: Can Heal Publications, 1992.

Hart IR, Harden RM, Walton HJ (eds) Newer Developments in Assessing Clinical Competence. Montreal: Heal Publications, 1986.

Hart IR, Harden RM (eds) Further Developments in Assessing Clinical Competence. Montreal: Can Heal Publications, 1987.

Lamberts H, Bouwer H, Mohrs J. Reason for encounter-, episode- and process-oriented standard output from the

Transition Project. Amsterdam: Department of General Practice University of Amsterdam, 1991.

Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ 1988;23:325-34.

Norman GR, Trott AD, Brooks LR, Smith EKM. Cognitive Differences in Clinical Reasoning Related to Postgraduate Training. Teach Learn Med 1994;6:114-20.

Patrick J. Training: research and practice. London: Academic Press, 1992:19-71.

Reznick RK, Smee S, Baumber JS. Guidelines for Estimating Real cost of an Objective Structured Clinical Examination. Acad Med 1993;68:513-517.

Stillman P, Swanson D, Smee S. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986; 105:762-71.

Stillman PL, Regan MB, Swanson DB. An Assessment of Clinical Skills of Fourth-Year Students at Four New England Medical Schools. Acad Med 1990;65:320-6.

Van der Vleuten CPM, Swanson DB. Assessment of skills with standardized patients: state of the art. Teach Learn Med 1990;2:58-76.

Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. Med Educ 1988;23:97-107.

Wooliscroft JO, TenHaken J, Smith J, Calhoun JG. Medical students' clinical self-assessments: comparisons with external measures of performance and the students' self-assessments of overall performance and effort. Acad Med 1993; 68:285-94.

Hoofdstuk 5

Performance-based assessment in continuing medical education for general practitioners: construct validity*

Summary

The use of performance-based assessment has been extended to postgraduate education and practising physicians, despite criticism on validity. While differences in expertise at this level are easily reflected in scores on a written test these differences are relatively small on performance-based tests. However, scores on written tests and performance-based tests of clinical competence generally show moderate correlations.

A study was designed to evaluate construct validity of a performance-based test for technical clinical skills in continuing medical education for general practitioners, and explore the correlation between performance and knowledge of specific skills.

A one-day skills training was given to 71 general practitioners, covering four different technical clinical skills. The effect of the training on performance was measured with a performance-based test using a randomized controlled trial design, while the effect on knowledge was measured with a written test administered one month before and directly after the training.

A training effect could be shown by the performance-based test for all four clinical skills. The written test also demonstrated a training effect for all but one skill. However correlations between scores on the written test and on the performance-based test were low for all skills.

It is concluded that construct validity of a performance-based test for technical clinical skills of general practitioners was demonstrated, while the knowledge test score showed to be a poor predictor of competence for specific technical skills.

Introduction

In measurement of clinical competence the use of direct observation of clinical performance under standardized conditions has become a popular assessment method, because it directly assesses behaviour considered relevant to clinical performance. The method has been extensively studied, providing general supportive evidence for validity and acceptable reliability (Van der Vleuten & Swanson 1990; Colliver & Williams 1993; Vu & Barrows 1994).

Performance-based testing has also been extended to postgraduate education (Stillman *et al.*1986; Cohen *et al.*1990; Joorabchi 1991; Grand'Maison *et al.* 1992) and assessment of practising physicians (Norman *et al.*1993; Rethans *et al.* 1991; Jansen *et al.* 1995). However,

---

* Published as: Jansen JJM, Scherpbier AJJA, Metz JCM, Grol RPTM, Van der Vleuten CPM, Rethans JJ. Performance-based assessment in continuing medical education for general practitioners. Med Educ 1996; 30: 339-44.

the use of this method to assess clinical competence at postgraduate level and among practising physicians has been criticised for lack of validity, because of the rigidity (Cox 1990) or trivialization (Norman *et al.* 1991) of the scoring methods used. These critics suggest that the assessment method based on checklists may be appropriate to assess basic history-taking and physical examination skills, but not in discriminating between different levels of expertise at graduate level and beyond. Nevertheless validation studies have shown (small) differences in mean scores between senior students and residents (Cohen *et al.* 1990; Joorabchi 1991; Brailovsky *et al.* 1995), and between junior and senior-levels within residency-training (Stillman *et al.* 1986; Petrusa *et al.* 1990). Few studies have included comparison of residents and practising physicians. In two experiments comparing residents in family medicine and practising physicians, no overall differences in score were found, although one study reported differences in subscores (Jansen *et al.* 1995; Brailovsky *et al.* 1995). This finding could be explained by the failure of the instrument to measure relevant differences in clinical competence as well as by the failure of the theory underlying the construct, i.e. practising physicians are more competent in the skills assessed in the test compared to residents (Crocker & Algina 1986).

One way to further evaluate construct validity is to assess the discriminating power of performance-based tests among groups of practising physicians with differences in competence. Norman *et al.* (1993) compared a criterion group of competent physicians, with self-referred physicians and physicians referred by the licensing body because of deficiencies, using multiple assessment methods, and found significant differences on the standardized patient-based test but not on the objective structured clinical examination.

Written tests can discriminate very well between different levels of competence at postgraduate level compared to performance-based tests (Swanson *et al.* 1987; Quattlebaum *et al.* 1989; Benson 1991; Norman *et al.* 1994), but have been criticised for lack of validity beyond recall of knowledge (Levine *et al.* 1970; Dixon 1978; Neufeld 1985). However, studies correlating results on written and performance-based test formats have found moderate to high true correlations (Van der Vleuten & Swanson 1990), providing supportive evidence for the assumption of a relation between knowledge and performance of clinical skills (Miller 1990). It has been argued that these high correlations are perhaps a result of memorizing the checklists used in the performance-based test (Norman *et al.* 1991; Van Luijk *et al.* 1990), but in a recent study among family physicians, not familiar with the content of checklists used, also a moderate correlation was found between scores on a written test and a performance-based test covering a broad domain of technical clinical skills (Jansen *et al.* 1995). In continuing medical education, however, courses focus on specific topics rather than on a broad domain, and it is not clear if the correlation between knowledge and performance is as high for specific skills.

An experimental study was designed to evaluate construct-validity of a performance-based test for technical clinical skills in continuing medical education of general practitioners, and compare results for the specific skills on the performance-based test with scores on a written

test of skills. Our research questions were:

1. Can the performance-based test discriminate between groups of practising physicians with different competence for specific technical clinical skills?
2. How accurately can the results for specific skills on the performance-based test be predicted by the scores on corresponding parts of the written test?

Methods

A 1-day training course in technical clinical skills was developed. The training focused on four topics: physical examination of the shoulder, injection techniques of the shoulder, cardiopulmonary resuscitation and intravenous cannulation. These topics were selected as having priority based on a survey among twenty general practitioners actively involved in CME throughout the country. Training was based on national clinical guidelines developed by professional bodies (Grol 1990). The training time was one hour for each topic, and each training was given in small groups (8-12 participants) by two experienced trainers with special interest in the subject concerned. It was assumed that such a training would result in a considerable improvement in competence.

Instruments/Materials

The effect of the skills training on performance was assessed by a performance-based test consisting of four OSCE-stations, covering the four topics addressed in the course. Checklists were used for scoring performance and for providing feedback, with criteria based on the national guidelines for general practice. The checklist for examination of the shoulder contained 36 items, for injection of the shoulder 20 items, for resuscitation 16 items, and 25 items for intravenous cannulation. Checklists were developed by a committee of general practitioners, reviewed by at least three faculty members and pilot-tested before the course. In addition to the checklist a ten point global rating scale was used as a general measure of performance.

In one station (shoulder examination) students with experience as standardized patients were used. They were trained for their role by a general practitioner experienced in the training of standardized patients in a two hour training session. In the other stations manikins (Resusci-Anni® CPR model, Limbs&Things® shoulder injection model, Syma® arm model) were used.

A total of 36 general practitioners (staff members from two departments of general practice) were involved as raters. One third of the encounters were double-rated to determine interrater reliability. Two weeks before the course the raters received a one hour training. To improve consensus, scoring was practised in the training-session and interrater differences were discussed.

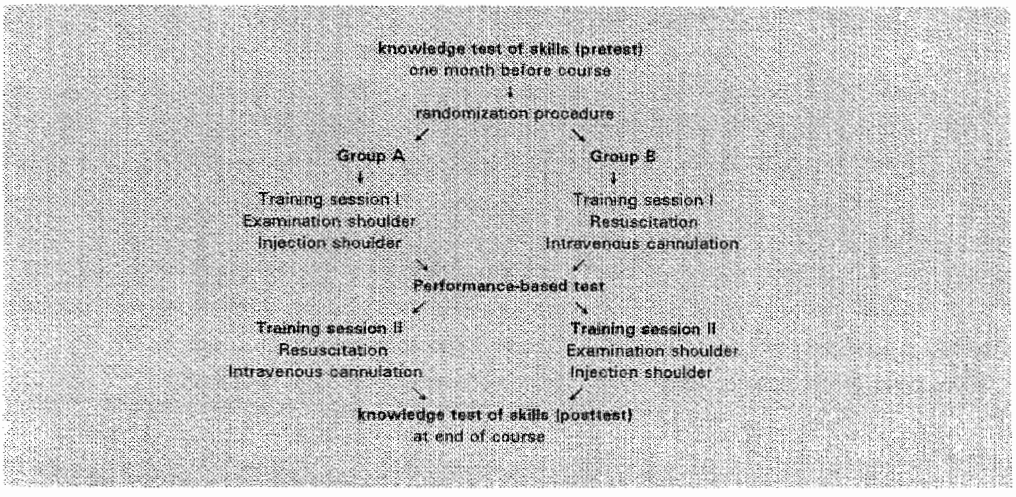
The effect of the skills training on knowledge of the participants was assessed by a written test, covering the content of the course. The 49 items consisted of statements with three answering options: true, false or question mark. The statements covered knowledge about the four technical clinical skills. The number of items for each topic was based on the number of relevant statements that could be constructed, resulting in 20 items about shoulder examination, 10 items

about shoulder injection, 13 items about resuscitation. Only six items about intravenous cannulation were included because it proved difficult to construct more meaningful questions about this technical skill.

Procedure

The course was announced in a mailing to general practitioners in the region. Participants (n=71) were divided at random into two groups. At the course one group (A, n=32) started with the training of shoulder examination and injection techniques, followed by the training on resuscitation and intravenous cannulation, while the other group (B, n=39) received the training in the opposite order (see figure 1). The performance-based test was administered between the two training sessions.

Figure 1 Design for training course and assessment sequence



Because of the randomized assignment of the participants the two groups could serve as each others controls for the different topics. As group A received the training on examination and injection of the shoulder before entering the performance-based test, while group B received this training after the test, the effect of this training could be evaluated by comparing the scores of both groups on the stations assessing examination and injection of the shoulder. The same comparison could be made for resuscitation and intravenous cannulation, were group A served as a control for group B. The participants received immediate feedback at each station on their performance by the rater using the checklist.

The knowledge test was mailed to all participants one month before the course and again administered directly after the training (pretest-posttest design). Participants only received feedback on their scores after the posttest.

Statistical analysis

The complete results on all four performance stations were available for 71 participants. As ten participants failed to return the written pretest, complete data on the written tests were available for 62 participants. Raw scores on the performance-based test and on the written test (number of correct items) were converted into a percentage score, and T-test was used to compare mean scores. Reliability of the knowledge test score was determined by calculating a Cronbach's alpha reliability coefficient (Cronbach *et al.* 1972) and for the performance-based test interrater reliability was assessed using intra-class correlation coefficients (Kramer & Feinstein 1981). Correlations between knowledge test score and performance-based test score were determined using Pearson product-moment coefficients (Welkowitz *et al.* 1982).

Results

Subjects

The 71 participants had a mean age of 41 years (range 30-55) and 10 years of experience (range 1-24) as family doctors. Most physicians (69%) worked full-time in their practice, with the remaining working 3-4 days (20%) or less (11%) in a practice. Mean practice size was 2500 patients (range 600-3600). Practice localizations were largely (sub)urban (41%) or small town (36%), and 23% were rural. Only 25% worked in a solo-practice, 41% in a duo-practice and the remaining 34% worked in a group practice or health centre. Compared to the population of Dutch general practitioners, there were more female doctors and part-timers among the participants, and fewer doctors working in a solo-practice, while age distribution, practice size and practice localization of the participants can be considered as representative. Doctors in group A (n=32) and group B (n=39) did not differ in characteristics, nor on the written test score prior to the course, suggesting that randomization had been successful.

Reliability

The interrater reliability coefficients for the checklist scores on the four stations of the performance-based test were: 0.97 for examination of the shoulder, 0.98 for injection of the shoulder, 0.93 for intravenous cannulation and 0.79 for resuscitation (the values based on the rating scale were respectively 0.88, 0.89, 0.75 and 0.70). These figures indicate that interobserver variability was minimal. The reliability coefficient for the written test was 0.72 for the pretest and 0.64 for the posttest.

Scores

Table 1 shows the results for the performance-based test for the checklist-score and rating scale. Before training, the mean scores on all stations revealed considerable deficiencies in performance, especially for the shoulder injection, while performance concerning resuscitation was relatively good. Based on the checklist score a significant improvement was found on all stations after training, with a mean increase in score of 24% (range 12-35%) of the maximum score, and smaller standard deviations in the group who had received training on three of four

topics, supportive for a training effect. The increase of the score on the resuscitation station was somewhat lower compared to the other stations. The rating scale scores mirrored closely the checklist scores, with ratings being only somewhat less stringent for pre-training performance on the shoulder stations.

Table 1 Checklist and rating scale scores on the performance-based test

	Checklist				Rating scale		
	n	mean	sd	T-test*	mean	sd	T-test*
Examination shoulder							
before training	39	51.5	15.8	p<0.001	65.9	13.1	p<0.001
after training	32	73.7	10.3		76.9	11.2	
Injection shoulder							
before training	39	38.7	20.6	p<0.001	50.9	16.3	p<0.001
after training	32	73.8	14.6		72.2	10.4	
Resuscitation							
before training	32	65.8	12.2	p<0.001	60.0	11.8	p<0.001
after training	39	78.0	12.3		75.4	10.9	
Intravenous cannulation							
before training	32	50.3	25.5	p<0.001	49.4	24.4	p<0.001
after training	39	77.9	15.0		76.4	14.4	

Note: all scores expressed as percentage of maximum score * T-test for difference before-after training

Table 2 Scores on the written test before and after training

	mean	sd	paired T-test
Examination shoulder (20 items)			
before training	66.3	15.8	p<0.001
after training	79.3	13.0	
Injection shoulder (10 items)			
before training	54.8	20.1	p<0.001
after training	77.7	13.0	
Resuscitation (13 items)			
before training	56.9	13.2	p<0.001
after training	67.6	13.8	
Intravenous cannulation (6 items)			
before training	75.5	22.3	p=0.264
after training	78.8	14.2	

note: all scores expressed as percentage of maximum score

Table 2 provides the scores on the written test a month before and directly after the training for the different topics. The scores showed significant improvement on all topics except for intravenous cannulation. The pretest score for intravenous cannulation was high, indicating that questions were probably relatively easy and limiting possibility of improvement.

Correlation

The scores on the checklists and the general ratings were correlated for all four stations, resulting in a correlation coefficient of 0.80 for examination of the shoulder, 0.87 for injection of the shoulder, 0.60 for resuscitation and 0.80 for intravenous cannulation. The checklist scores on the performance-based test were correlated with the pretest scores and posttest scores on the knowledge test. The scores on the performance-based stations for participants before the training were matched with their scores on the corresponding parts of the written pretest, while for the scores on the stations after the training the corresponding parts of the written posttest were used. The results are presented in table 3. Correlations between scores on the knowledge test and the performance-based test are variable, decreasing from significant to non significant after training for 'injection of the shoulder', while increasing to significant ($p < 0.05$) for 'examination of the shoulder' and 'resuscitation'. Correlation of the general rating with the written tests resulted in comparable figures.

Table 3 Correlations of the performance-based test scores (checklist and rating scale) with the knowledge test scores

	checklist	rating scale
examination shoulder		
pretest score	0.20	0.28
posttest score	0.43*	0.23
injection shoulder		
pretest score	0.35*	0.30
posttest score	-0.20	0.03
resuscitation		
pretest score	0.14	-0.20
posttest score	0.35*	0.01
intravenous cannulation		
pretest score	0.24	0.24
posttest score	-0.05	-0.20

* $p<0.05$

Discussion

A considerable training effect was demonstrated on the performance-based test (both on the checklist and on the general rating scale) for all four clinical skills in a short hands-on skills training in small groups for practising family physicians. These results suggest that a performance-based assessment method can indeed discriminate between different levels of proficiency among practising physicians, which provides support for construct validity. Other recent studies have demonstrated similar results for different technical clinical skills (Nyquist *et al.* 1994; Carney *et al.* 1995). Interrater reliability was high as has been reported in other studies concerning clinical skills (Wakefield 1985), with rating scales having a somewhat lower reliability (Van Luijk & Van der Vleuten 1992).

The knowledge test score also improved for all but one skill as a result of the training. The knowledge test failed to demonstrate a training effect for intravenous cannulation, while performance did improve by more than 25%. A likely explanation is that questions in the knowledge test were too easy, so discriminating power was lost.

Correlations between checklist scores and general ratings were high for all stations, except resuscitation, which showed a moderate correlation. This could indicate that some relevant performance aspects were not well covered by the checklist. For the other three stations the high correlations with the general ratings are supportive for content validity of the checklist, since the raters were experienced general practitioners, and therefore were considered experts in the evaluation of performance of their peers. These results indicate that both rating scales and checklists seem appropriate measurement tools in assessment of performance of technical clinical skills of general practitioners.

Correlations between scores on the written test and the performance-based test were variable but low. Even when leaving intravenous cannulation out of consideration because of the above mentioned problems, knowledge of a skill was not a reliable predictor of proficiency for that specific technical clinical skill, as knowledge predicted only a very small part of the variance on the performance-based test for the different skills. The low reliability of the written test used, may have had a negative influence on the correlations. However, the content of each specific skill puts a limit to the number of meaningful items from which a written test can be sampled, contrary to assessment of clinical competence as a general construct, where the domain from which items for test construction can be sampled is very large. Correction for unreliability was therefore not considered appropriate. The results are consistent with an earlier study (Vu & Barrows 1990). Although scores on knowledge tests and performance-based tests can have a high correlation when generalized over a broad domain (Van der Vleuten *et al.* 1988; Newble & Swanson 1988; Jansen *et al.* 1995), this relation is not necessarily replicated for specific skills.

In conclusion, while both performance-based test and written test were able to demonstrate a training effect, they apparently measured different things: performance ('shows how') and knowledge ('knows'), applying the terminology of Miller (1990). Knowledge, perhaps useful as a predictor of performance when generalized over a broad domain, resulted to be a poor

predictor of performance for specific technical skills. For assessment of mastery of specific technical clinical skills a performance-based test is preferably used, and both checklists and rating scales seem suitable.

References

Benson JA. Certification and Recertification: one approach to professional accountability. Ann Intern Med 1991;114:238-42.

Brailovsky CA, Grand'Maison P, Lescop J. Construct validity of the objective structured clinical examination used in the Quebec licensing examination. In: Rothman AI, Cohen R (Eds). Proceedings of the sixth Ottawa conference on medical education. University of Toronto Bookstore, Toronto, 1995:373-4.

Carney PA, Dietrich AJ, Freeman DH, Mott LA. A standardized-patient assessment of a continuing medical education program to improve physicians' cancer-control clinical skills. Acad Med 1995;70:52-8.

Cohen R, Reznick RK, Taylor BR, Provan J, Rothman A. Reliability and validity of the objective structured clinical examination assessing surgical residents. Am J Surg 1990;160:302-5.

Colliver JA, Williams RG. Technical issues: test application. Acad Med 1993;68:454-60.

Cox K. No Oscar for OSCA. Med Educ 1990;24:540-5.

Crocker L, Algina J. Introduction to classical and modern test theory. Harcourt Brace Jovanovich, Orlando, Florida, 1986:232.

Dixon J. Evaluation criteria in studies of continuing education in the health professions: a critical review and a suggested strategy. Eval Health Professions 1978;1:47-65.

Fisher EW, Pfleiderer AG. Assessment of the otoscopic skills of general practitioners and medical students: is there room for improvement? Br J Gen Pr 1992;42:65-7.

Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective structured clinical examination for licensing family physicians. Can Med Assoc J 1992;146:1735-40.

Grol RPTM. National standard setting for quality of care in general practice: attitudes of general practitioners and response to a set of standards. Br J Gen Pr 1990;40:361-4.

Jansen JJM, Tan LHC, Van der Vleuten CPM, Van Luijk SJ, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners. Med Educ 1995;29:247-53.

Joorabchi B. Objective structured clinical examination in a pediatric residency program. Am J Dis Child 1991;145:757-62.

Kopelow ML, Schnabl GK, Hassard TH, Tamblyn RM, Klass DJ, Beazley G, Hechter F, Grott M. Assessing practising physicians in two settings using standardized patients. Acad Med 1992;67:S19-21.

Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981;20:111-23.

Levine HG, McGuire CH, Nattress LW. The validity of multiple choice achievement tests as measures of competence in medicine. Am Educ Res J 1990;1:69-83.

Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-7.

Neufeld VR. Written examinations. In: Neufeld VR & Norman GR (eds). Assessing clinical competence. Springer, New York, 1985:94-118.

Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ 1988;23:325-34.

Norman GR, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Med Educ 1991;25:119-26.

Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. JAMA 1993;270:1046-51.

Norman GR, Trott AD, Brooks LR, Smith EKM. Cognitive differences in clinical reasoning related to postgraduate training. Teach Learn Med 1994;6:114-20.

Nyquist JG, Naylor AJ, Woodward-Lopez G, Dixon S. Use of performance-based assessment to evaluate the impact of a skill-oriented continuing education program. Acad Med 1994;69:S51-3.

Quattlebaum TG, Darden PM, Sperry JB. In-training examinations as predictors of resident clinical performance. Pediatrics 1989;84:165-72.

Petrusa E, Blackwell T, Ainsworth M. Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. Arch Intern Med 1990;150:573-7.

Rethans JJ, Sturmans F, Drop R, Van der Vleuten C, Hobus P. Does competence of general practitioners predict their performance. Br Med J 1991;303:1377-85.

Stillman PL, Swanson DB, Smee S, Stillman AE, Ebert TH, Emmel VE, et al.. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986;105:762-71.

Swanson DB, Norcini J, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. Assessment Eval Higher Educ 1987;12:220-46.

Van der Vleuten CPM, Van Luijk SJ, Beckers HJM. A written test as an alternative to performance testing. Med Educ 1988;23:97-107.

Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients; state of the art. Teach Learn Med 1990;2:58-76.

Van Luijk SJ, Van der Vleuten CPM, Van Schelven SM. The relationship between content and psychometric characteristics in performance-based tests. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds) Teaching and Assessing Clinical Competence. Boekwerk, Groningen, 1990:202-7.

Van Luijk SJ, Van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM (eds) Current Development in Assessing Clinical Competence. CanHeal, Montreal, 1992:357-62.

Vu NV, Barrows HS. Validity and accuracy of performance and written evaluations in assessing history and physical examination skills. In W Bender, RJ Hiemstra, AJJA Scherpbier, RP Zwierstra (eds), Teaching and Assessing Clinical Competence. Boekwerk, Groningen, 1990:283-7.

Vu NV, Barrows HS. Use of standardized patients in clinical assessments: recent developments and measurement findings. Educ Res 1994;23:23-30.

Wakefield J. Direct Observation. In: Neufeld VR & Norman GR (eds). Assessing clinical competence. Springer, New York, 1985:51-70.

Welkowitz J, Ewen RB, Cohen J. Introductory statistics for the behavioral sciences. Harcourt Brace Jovanovich, Orlando, Florida, 1982:177.

Hoofdstuk 6

Evaluation of cardiopulmonary resuscitation skills of general practitioners using different scoring methods*

Summary

In this study we evaluated the practical performance of 70 general practitioners in cardiopulmonary resuscitation (CPR) before and after instruction and compared checklist-based scores to mechanical recording scores in order to investigate which scoring method is preferable.
Both checklist and recording strip based scores showed significant improvement after instruction, but only 37% were judged proficient according to the American Heart Association standards (checklist scoring), and 47% according to the recording print based scoring system, while raters judged 97% as satisfactory by general impression. Interrater reliability was highest for the recording print (0.97) and lower for the checklist (0.79), especially for CPR-performance (0.56). Comparison of checklist and recording print showed that the checklist was specific but not very sensitive in identifying poor performance for cardiac compression rate, since observers overestimated performance. The correlation for CPR-performance between checklist score and recording strip score was low (0.45), indicating that candidates were ranked differently. The correlation between diagnosis and performance score was low for checklist as well as recording print (0.22), indicating that the score on diagnosis was a poor predictor for the score on performance of CPR.
These results support the use of the recording manikin as compared with the use of a checklist for formative evaluation of basic life support skills. However, as proficiency in diagnosis and performance in CPR are poorly correlated, assessment of diagnosis using a checklist must be included. Therefore we strongly recommend the combination of assessment by observers using a checklist for diagnostic procedures and the recording strip of the manikin for performance of CPR, as employed in most evaluation schemes.

Introduction

Cardiopulmonary arrest is a frequent cause of death in the developed world, with approximately two-thirds of the deaths occurring outside hospital [1]. Research evidence suggests that rapid initiation as well as correct technique of cardiopulmonary resuscitation (CPR) are essential links in the 'chain of survival' [2,3]. Since the majority of sudden deaths occur in the community, many lives could possibly be saved if adequate CPR skills were present throughout the

* Published as: Jansen JJM, Berden HJJM, Van der Vleuten CPM, Grol RPTM, Rethans JJ, Verhoeff CPM. Evaluation of cardiopulmonary resuscitation skills of general practitioners using different scoring methods. Resuscitation 1997; 34: 35-41.

community. General practitioners are confronted each year with 5-10 patients suffering from acute myocardial infarction [4,5]. The reported risk of cardiac arrest before reaching hospital varies from approximately 5% [6] up to 25% [7]. In a recent survey in the Netherlands general practitioners reported a mean performance of 2,0 CPR-attempts per year [8]. Various studies have shown considerable deterioration in CPR-skills among physicians, who had successfully completed prior courses in CPR [9-12], indicating that proficiency in these skills is not maintained.

For evaluation of competence in basic life support, checklists covering criteria of adequate performance are used [13] as well as recording strips of manikins [14]. In most research a combination of these methods is used [12,15-17], with checklist-based scoring for diagnostic procedures and the recording strip of the manikin for compression and ventilation procedures. The use of recording manikins permits assessment of outcome criteria (e.g. breathing volume and thorax impression depth) and some aspects of process, while checklists tend to concentrate on process criteria (e.g. how the ventilation procedure is performed, and position of shoulders and hands of the resuscitator during thorax compression), which are considered to be relevant for outcome. Moreover checklists can be used for scoring of the diagnostic assessment of the victim, which cannot be assessed by the recording manikin.

Only limited research has addressed comparison of checklist and recording strip as evaluation methods for CPR. Two authors reported comparisons between checklist-based scores and mechanical recording based scores [17,18], and concluded that checklist-based scores overestimated competence.

In this study we evaluated the practical performance of general practitioners in cardiopulmonary resuscitation before and after instruction and compared checklist-based scores to mechanical recording scores to investigate which scoring method is preferable.

Materials and methods

Seventy-one general practitioners participated in a continuing medical education course with basic cardiopulmonary resuscitation as one of the topics. An account of this course has been published elsewhere [19]. The training time for CPR was one hour and training was given in small groups (8-12 participants) by two experienced CPR-trainers. Participants were randomly divided into two groups, one was evaluated before instruction and one was evaluated after instruction.

A checklist [20] was used for evaluation based on the guidelines of the Dutch Heart Association [21], comparable to the guidelines of the American Heart Association [22,23] and the European Resuscitation Council [24], except for the sequence used to initiate CPR. This checklist contained 16 items and included criteria for diagnostic assessment of unresponsiveness, circulation and airway in the correct sequence and speed (6 items) and correct sequence and performance of cardiopulmonary resuscitation procedures (10 items) (see table 1). Criteria for cardiac compression included correct placement of hands and position during cardiac compression, compression rate (80-100 per minute) and ratio for compression and ventilation (15:2).

Table 1 Checklist cardiopulmonary resuscitation (according guidelines Dutch Heart Association) [20,21]

	not or incorrectly performed	correctly performed
Diagnostic procedures		
1. Assessment of unresponsiveness		
- calls loudly on the victim	☐	☐
- gives strong pain-stimulus	☐	☐
2. Assessment of circulation		
- checks unilaterally for carotid pulse >4 seconds	☐	☐
3. Assessment of airway		
- checks if airway is free and assesses breathing	☐	☐
4. Diagnostic procedures are performed in correct sequence	☐	☐
5. Concludes diagnosis within 30 seconds	☐	☐
Performance of CPR		
6. Starts with chest compressions	☐	☐
7. Proper compression position		
- Correct position of shoulders	☐	☐
- Proper handplacement	☐	☐
8. Adequate compression technique		
- Maintains chest compression rate of 80-100 per minute	☐	☐
- Performs cycles of 15 compressions and two ventilations	☐	☐
9. Adequate ventilation technique		
- performs correct head tilt chin lift manoeuvre	☐	☐
- mouth fully covers mouth patient	☐	☐
- prevents air escaping from nose patient	☐	☐
- watches for chest movements during insufflation	☐	☐
10. Chest rises and falls during ventilation	☐	☐
General impression (1-10):		

Criteria for ventilation included correct head tilt-chin lift manoeuvre, prevention of air escape during ventilation and observing for chest rise and fall. Scoring of the checklist criteria allowed for marking adequate or inadequate performance. After scoring the separate criteria, raters were also requested to provide a general impression of CPR-proficiency on a ten-point rating scale.

The performance of cardiopulmonary resuscitation procedures was also assessed by the structured use of the recording strip of a resuscitation manikin (Laerdal Recording ResusciAnne type 20.00.10) as described by Berden et al. [14]. This scoring system includes criteria for placement of the hands, compression rate and depth, compression/relaxation ratio, and breathing volume and interval (see table 2).

After receiving standardized instruction participants were rated while performing single-rescuer cardiopulmonary resuscitation during two minutes on a ResusciAnne recording manikin. Then feedback on performance was provided by the rater, based on the checklist rating and the recording printout strip. One third of the encounters was double-rated to determine interrater reliability of the checklist-based score and the general impression rating. Raters were general practitioners recruited from the staff of two university departments of general practice and had

Table 2 Recording strip and scoring system cardiopulmonary resuscitation (Berden et al 1988)

parameter	value	penaltypoints
placement of hands	right	0
	wrong	5
compression rate (per minute)	80-100	0
	100-120	5
	120-140 or 60-80	10
	>140 or 40-60	15
	<40	20
compression depth (mm)	38-52	0
	30-38 or 52-60	5
	22-30 or >60	10
	<22	20
compression /relaxation ratio	0.6-1.4	0
	<0.6 or >1.4	10
breathing volume (liters)	0.8-1.2	0
	1.2-1.5	5
	0.4-0.8 or 1.5-2	10
	>2.0	15
	<0.4	20
breathing interval (seconds)	<4	0
	4-6	5
	6-8	10
	8-10	15
	>10	20

no specific experience as CPR-trainers (the two CPR-trainers were not included as raters). Two weeks before the course the raters received one hour of instruction to practice scoring and discuss interrater differences, the aim being to achieve consensus. The recording strips were scored after the course by the first and second author, and half of the strips were double-rated to determine interrater reliability.

Data management and statistical analysis

Complete scores on checklist and recording strip were available for 70 participants, as from one candidate no recording strip was available due to malfunctioning of the manikin. A general impression rating was available for 64 participants, and was missing for six participants. Raw scores on the checklist (maximum score 16 points), general impression rating (maximum 10 points) and recording strip (maximum 95 penalty points) were converted into a percentage score, after penalty scores on the recording strip were reversed to bonus scores. The Mann-Whitney two-tailed test was used to compare mean scores before and after training. Pass-not yet passed decisions were based on the performance of CPR using the standard of the American

Heart Association [22] for the checklist (i.e. no errors allowed) and the standard set by Berden [14] for the recording strip, allowing a maximum of 15 penalty points. For the general impression rating a score of 6 or more was considered a pass-score. The methods were compared with regard to reliability using intraclass correlation coefficients [25]. Accuracy of observer assessment based on checklist criteria also covered by the recording strip was measured calculating sensitivity and specificity indexes [26], with the recording strip serving as gold standard. Consistency in ranking between the different methods was measured with Spearman's rank correlation coefficient.

Results

Scores

In table 3 the mean scores (SD) are given for the checklist, rating scale and recording print for the group before (n=32) and after instruction (n=38). Mean scores were lowest for diagnosis and showed no significant improvement after instruction. Scores on performance were higher for the checklist compared to the recording strip, and showed improvement on both scoring methods. However, this difference was not statistically significant for checklist-based ventilation. Finally, total checklist-based score and rating scale showed difference in score before and after instruction. Applying the standard of the American Heart Association to the checklist resulted in 5/32 (15%) participants with adequate CPR-performance before instruction and 14/38 (37%) after instruction. For the scoring system based on the recording print the figures rose from 6/32 (18%) before to 18/38 (47%) after instruction. Based on the general impression rating, pass-scores were 17/28 (61%) before and 35/36 (97%) after instruction.

Table 3 Checklist and recording print scores cardiopulmonary resuscitation: results before and after training*

	Checklist (Mean (SD))		Recording print (Mean (SD))	
	Before	After	Before	After
Diagnosis	52.6 (27.5)	61.0 (21.7)		
Performance	75.0 (15.5)	87.1 (11.4) [b]	68.1 (11.6)	78.0 (10.6) [b]
Cardiac compression	64.4 (22.0)	86.3 (18.7) [b]	73.3 (13.5)	82.8 (13.8) [a]
Ventilation	83.1 (19.7)	90.0 (11.2)	60.9 (21.7)	72.4 (18.6) [a]
Total score	65.8 (12.2)	78.0 (12.3) [b]		
General impression (rating scale)	60.0 (11.9)	75.4 (10.8) [b]		

* Scores are presented as percentage of maximum scores
[a] Mann-Whitney $p<0.05$ for difference before-after
[b] Mann-Whitney $p<0.001$ for difference before-after

Interrater reliability

The interrater reliabilities for the overall scores on the different assessment methods and checklist subscores are shown in table 4. The reliability of the score based on the recording strip was highest, while the general impression rating showed the lowest reliability. For the checklist interrater reliability of the total score and diagnosis was much higher than for performance.

Table 4 Interrater reliability for general impression, checklist and recording strip CPR

	interrater reliability*
General impression	0.70
Total score checklist	0.79
Subscore diagnosis (item 1-5)	0.77
Subscore performance (item 7-10)	0.66
Recording strip	0.97

* Intraclass correlation coefficient

Observer accuracy

It was possible to evaluate accuracy of observers on two criteria, compression rate and ventilation volume, which were covered both by the checklist and recording strip. The recording strip scores were dichotomized according to checklist criteria: for compression rate cutoff points were 80 and 100 compressions per minute, while for the ventilation volume the cutoff point was 0.8 litre, considered equivalent to the mimimum volume necessary to make the chest rise [23]. As shown (table 5) observers judged a higher number of participants as performing adequately for compression rate and ventilation volume, compared to the recording results. The sensitivity and specificity indexes reveal that observers were specific but not very sensitive in identifying poor performance on the two criteria.

Table 5 Accuracy of checklist marking for compression rate and ventilation volume*

	Compression rate 80-100/min		Ventilation volume >0.8 L	
	Yes	No	Yes	No
Recording strip*	32	38	52	18
Checklist*	42	28	52	8
Agreement*	26	22	51	7
Sensitivity	0.81	0.58	0.98	0.39
Specificity	0.58	0.81	0.39	0.98

* Recording strip as golden standard

Consistency in ranking

The Spearman rank correlation coefficients between the different scores are shown in table 6. The correlation between checklist score and general impression was moderate (0.67). The correlation for CPR-performance between the checklist score and recording strip score was low (0.45), indicating that the two methods apparently ranked participants quite differently. Also the correlation between the diagnosis score and performance score was low (0.22) for the checklist as well as for the recording strip.

Table 6 Correlations* between different methods

		Checklist		
	General impression	Total score	Diagnosis (item 1-5)	Performance (item 7-10)
Recording strip total score	0.45	0.45	0.22	0.46
Checklist performance score	0.67	0.71	0.22	
Checklist diagnosis score	0.25	0.79		
Checklist total score	0.61			

* Spearman rank-order coefficient

Discussion

General practitioners showed considerable deficiencies in basic cardiopulmonary resuscitation skills. This confirms results of earlier studies among different health professionals [9-12]. A one- hour refresher course improved scores but was not enough for all participants to acquire an adequate level of performance according to the scoring system based on the recording strip or criteria of the American Heart Association. However, the general impression of the raters was much more favourable. As raters were general practitioners they could have been reluctant to judge their peers as performing unsatisfactorily. On the other hand pass-not yet passed decisions based on the standard format of the American Heart Association or the recording strip may be unnecessarily stringent concerning CPR-performance procedures. For early activation of the emergency medical service and rapid initiation of basic life support the effect on outcome is well demonstrated [2,3,27-29], while evidence for the effect on survival of variability in performance of cardiac compression or ventilation is not substantial. Lund [2] demonstrated a negative effect on survival of gross omissions in CPR technique (e.g. performing ventilation without cardiac compression). In other investigations no relation was found between level of CPR skills and patient outcome [30]. The high standards, as used in this study, perhaps have more significance as an educational goal of excellence and are not necessarily critical for survival. The formative value of CPR assessment, which allows providing of immediate detailed feedback to trainees, should therefore be emphasized rather than its summative value, in order to avoid possible discouraging effects on motivation to perform CPR [31].

The comparison of a checklist-based and a recording strip-based scoring system revealed considerable differences between these methods. The interrater reliability for the checklist was comparable to those reported in the literature for technical clinical skills [32,33], but was lower compared to the recording strip, as recording strip-scoring allowed less observer error. Nevertheless interrater reliability for the diagnostic procedures was very acceptable, indicating that observers agreed strongly about scoring in this part of the checklist. This provides support for the use of a checklist for scoring of the diagnostic procedures.

The interrater reliability was considerably lower for the performance of CPR (cardiac compression and ventilation), indicating that perhaps observation criteria for behaviour during cardiac compression and ventilation were less clear or procedures themselves were more difficult to observe. Moreover, accuracy of checklist scoring for compression rate, using the recording strip as gold standard, was low. Although raters were specific in identifying poor performance, they were not very sensitive, as they tended to overestimate correct performance. Others have reported similar results [17,18]. For ventilation volume the difference between checklist scoring and recording strip may be a consequence of criteria used, because apparently also volumes lower than 0.8 litres will make the chest of the recording manikin rise [35]. Recently stronger emphasis on observation of chestrise as criterium for adequate ventilation has been recommended [36], so recording strip criteria used in this study were perhaps less valid compared to checklist criteria. Finally, the correlation between checklist score and recording strip score was rather low, indicating that candidates were ranked differently according to their scores in the two methods, as has been reported earlier [17]. These results support the superiority of the recording manikin print as compared with the checklist to evaluate performance of cardiac compression, while the study does not allow conclusions concerning preferable method for ventilation volume.

The correlation between rating scale and checklist was moderate, and higher than between rating scale and recording strip. This may have been caused by a 'halo effect' on the checklist (i.e. the raters' general impression of performance influenced the scoring of the separate criteria)[34]. The rather low correlation between the checklist and recording strip score for performance of CPR indicates that 'process'-oriented and 'outcome'-oriented assessment ranked resuscitators differently. Therefore, if feasible, a recording strip should be used to evaluate performance of CPR. Within the checklist correlation between diagnosis and performance score was low, as well as correlation between diagnosis and recording strip, indicating that the score on diagnosis is a poor predictor for the score in performance of CPR and vice versa. This has important implications for assessment of proficiency in CPR, because proficiency in diagnostic procedures should not apparently be taken for granted in individuals who demonstrate proficiency in performance of CPR.

Therefore we strongly recommend the combination of assessment by raters using a checklist for diagnostic procedures and the recording strip of the manikin for performance of CPR, as employed in most evaluation schemes.

Acknowledgements

The authors thank the SVUH (National institute for evaluation of vocational training) for permission to use the checklist, Jeroen Pielage for statistical support, and members of the Skillslab of the University of Limburg (head of department: Albert Scherpbier) and the Clinical Training Centre in Nijmegen (head of department: Jaap Metz) for their contribution. Gillian Hull of the General Practitioner Writers Association is thanked for polishing the English. This study was financially supported by a grant from the Dutch Ministry of Health.

References

1. Gillum RF, Folsom A, Luepker RW et al. Sudden deaths and acute myocardial infarction in a metropolitan area, 1979-1980. N Engl J Med 1983;316:1353-7.

2. Lund I, Skulberg A. Cardiopulmonary resuscitation by lay people. Lancet 1967;ii:702-4.

3. Eisenberg MS, Bergner L, Hallstrom A. Cardiac resuscitation in the community. JAMA 1979;241:1905-07.

4. Hodgkin K. Towards earlier diagnosis in primary care. Edinburgh: Churchill Livingstone,1978:252.

5. Lamberts H, Brouwer H, Mohrs J. Reason for encounter- episode- and process-oriented standard output from the Transition Project. Amsterdam: University of Amsterdam, 1991.

6. Pai GR, Haites NE, Rawles LM. One thousand heart attacks in Grampian: the place of cardioplumonary resuscitation in general practice. Br Med J 1987;294:352-4.

7. American Heart Association, Emergency Cardiac Care Committee and Subcommittees. Guidelines for cardiopulmonary resuscitation and emergency cardiac care, I: introduction. JAMA 1992;268:2172-83.

8. Dogger CA, Teijink JAW, den Blanken MM, Patka P, Haarman HJT. De rol van de huisarts in de preklinische spoedeisende hulpverlening (the role of the general practitioner in prehospital emergency care). Med Contact 1992; 47:205-7.

9. Stross JK. Maintaining competency in advanced cardiac life support skills. JAMA 1983;249:3339-41.

10. Kaye W, Mancini ME. Retention of cardiopulmonary resuscitation skills by physicians, registered nurses, and the general public. Crit Care Med 1986;14:620-2.

11. Berden HJJM, Willems FF, Ten Have FTM, et al. De primaire reanimatievaardigheid van de huisarts (Basic life support skills of general practitioners) Ned Tijdschr Geneesk 1988;132:1797-1801.

12. Seraj MA, Naguib M. Cardiopulmonary resuscitation skills of medical professionals. Resuscitation 1990;20:31-9.

13. Lowenstein SR, Hansborough JF, Libby LS, et al. Cardiopulmonary resuscitation by medical and surgical house-officers. Lancet 1981;ii:679-81.

14. Berden HJJM, Pijls NHJ, Willems FF, et al. A scoring system for basic cardiac life support skills in training situations. Resuscitation 1992;23:21-31.

15. Ramirez AG, Weaver FJ, Raizner AE, et al. The efficacy of lay CPR instruction: an evaluation. Am J Public Health 1977;67:1093-5.

16. Fossel M, Kiskaddon RT, Sternbach GL. Retention of cardiopulmonary resuscitation skills by medical students. J Med Educ 1983;58:568-75.

17. Kaye W, Mancini ME, Rallis SF et al. Can better basic and advanced cardiac life support improve outcome from cardiac arrest? Crit Care Med 1985;13:916-20.

18. Van Kalmthout PM, Speth PAJ, Rutten JR et al. Evaluation of lay skills in cardiopulmonary resuscitation. Brit Heart J 1985;53:562-6.

19. Jansen JJM, Scherpbier AJJA, Metz JCM et al. Performance-based assessment in continuing medical education for general practitioners: construct validity. Med Educ 1996;30:339-44.

20. Werkgroep Vaardigheden SVUH. Reanimatie zonder hulp(middelen) (Single-rescuer Resuscitation). Utrecht: SVUH, 1993.

21. Van Drenth J. Wanneer elke seconde telt (When every second counts). Den Haag: Nederlandse Hartstichting, 1989.

22. Instructor's Manual for Basic Life Support. Dallas, Texas: American Heart Association, 1985.

23. American Heart Association, Emergency Cardiac Care Committee and Subcommittees. Guidelines for cardiopulmonary resuscitation and emergency cardiac care, II: Adult Basic Life Support. JAMA 1992;268:2184-98.

24. Basic Life Support Working Party of the European Reuscitation Council. Guidelines for basic life support. Resuscitation 1992;24:103-10.

25. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981; 29: 111-23.

26. Inglefinger JA, Mosteller F, Thibodeau LA et al. Biostatistics in clinical medicine. New York: Macmillan, 1983: 4-10.

27. Copley DP, Mantle JA, Rogers WJ et al. Improved outcome for prehospital cardiopulmonary collapse with resuscitation by bystanders. Circulation 1977;56:901-5.

28. Cummins RO, Eisenberg MS, Hallstrom AP et al. Survival of out-of-hospital cardiac arrest with early initiation of cardiopulmonary resuscitation. Am J Emerg Med 1985;3:114-9.

29. Hallstrom AP, Cobb LA, Swain M et al. Predictors of hospital mortality after out-of-hospital cardiopulmonary resuscitation. Crit Care Med 1985;13:927-9.

30. Tweed WA, Wilson E. Is CPR on the right track? Can Med Assoc J 1984;131:429-33.

31. Sigsbee M, Geden EA. Effects of anxiety on family members of patients with cardiac disease learning cardiopulmonary resuscitation. Heart Lung 1990;19:662-5.

32. Van der Vleuten CPM & Swanson DB. Assessment of clinical skills with standardized patients; state of the art. Teach Learn Med 1990;2:58-76.

33. Jansen JJM, Tan LHC, Van der Vleuten CPM et al. Assessment of competence in technical clinical skills of general practitioners. Med Educ 1995;29:247-53.

34. Streiner DL. Global rating scales. In: Neufeld VR, Norman GR, editors. Assessing Clinical Competence. New York; Springer, 1985:119-41.

35. Laerdal Company; personal communication.

36. Vliet J van. Consensus Basale Reanimatie van de Nederlandse Reanimatie Raad (Consensus Basic Life Support of the Dutch Resuscitation Council) Ned Tijdschr Geneeskd 1996;140:596-9.

Hoofdstuk 7

Effect of a short skills training course on competence and performance in general practice*

Summary

Background. Short focused training sessions for technical clinical skills are popular among general practitioners, but research-evidence with regard to its effect on performance in practice is not conclusive.

Aim. Evaluation of the efficacy of a short course of technical clinical skills to change performance in general practice.

Method. Subjects were self-selected general practitioners (n=59), who were unaware of the study design. They were assigned to the intervention group (n=31) or control group (n=28) according to their preference for date of course. The course covered four different technical clinical skills (shoulder injection technique, PAP-smear, laboratory examination of fluor vaginalis, opthalmoscopic control in diabetes mellitus). Main outcome measures used were pre- and post-training scores on a knowledge test (60 multiple choice items), and pre- and post-training performance of procedures in practice using a log-diary covering 20 days.

Results. Competence as measured with the knowledge test improved significantly as a result of the training and skills test scores were satisfactory after training. A significant effect on performance in practice was found for PAP-smear and shoulder injection technique, whereas no effect could be demonstrated for examination of fluor vaginalis and opthalmoscopic control in diabetes mellitus.

Conclusion. A good degree of competence is a necessary but not always a sufficient condition for a physician to change his performance in practice. While for some skills training seems adequate to bring about desired changes, for other skills more complex interventions are needed.

Introduction

Physicians in general practice perform many different diagnostic and therapeutic procedures[1,2], amounting in the Netherlands to more than 4000 performed procedures per 1000 patients per year[3]. Together with counseling and prescribing, technical clinical skills constitute the core of the work of general practitioners. While the importance of competence in technical skills of general practitioners is acknowledged, surveys have indicated that undergraduate and graduate

---

* Submitted as: JJM Jansen, Grol RPTM, Van der Vleuten CPM, Scherpbier AJJA, Crebolder HFJM, Rethans JJ. Effect of a short skills training course on competence and performance in general practice

training programs are not covering all relevant skills[2,4,5]. Consequently general practitioners entering practice may not always be sufficiently prepared to perform these skills. Moreover, skills acquired may deteriorate because of insufficient practice or innovations in general practice requiring the acquisition of new techniques[6]. All these factors may contribute to deficiencies of competence in technical clinical skills of practising physicians. Research evaluating the competence of practising general practitioners has indeed provided evidence of existing deficiencies in skills such as clinical breast examination[7], resuscitation[8], ophtalmoscopy[9], examination of shoulder[10] and otoscopic examination[11]. These deficiencies in skills may affect the quality of care provided, because of missed diagnosis, inadequate treatment or unnecessary referral[7-9,11,12].

One way to remedy these deficiencies is continuing medical education. The efficacy of continuing medical education has been extensively evaluated[13-17]. From these evaluations it is evident that in general CME works well in improving physician competence and generally less well in changing practice performance. Apparently improved competence does not necessarily result in changes of performance, because practical problems and organisational or social barriers may limit the application of what was learned[18,19].

Other interventions aimed at changing performance in practice, like audit and feedback programs, are more effective[16]. More complex strategies, using combinations of methods, result in more consistent and substantial effects[13,15,17]. However, the disadvantages of these complex strategies are often high cost and sophisticated organisational requirements, limiting feasibility beyond research-settings. Unfortunately, still little is known about precisely what elements work and why[20].

Short focused training sessions of technical clinical skills for general practitioners have been introduced in the Netherlands[10], developed according to educational principles as outlined by Stein[21]. The courses are short interactive training sessions, in small groups, with hands-on practicing of skills. They are popular among general practitioners, but little is known about their effects. In this study we assessed the efficacy of such short courses of technical clinical skills aimed at changing performance in general practice.

Methods

Subjects

General practitioners from the south-eastern region of the Netherlands (n=800) were mailed with an invitation to participate in an interactive hands-on training course on four different technical clinical skills as part of an experiment investigating transfer of skills from training to the practice environment. Those who agreed to participate knew they were participating in an experiment, but were not informed about the design of the study. Acceptance was determined by order of registration to a maximum of 32 participants per course. Participants were divided into two groups, according to their preference for the timing of the course, with group A (intervention) receiving the course three months earlier than group B (control).

Materials

The course covered four different skills: injection technique of the shoulder, ophtalmoscopic control in diabetes, PAP-smear and laboratory examination of fluor vaginalis. These skills were selected because competence in the procedures was known to be amenable to improvement[9,22-24]. For each skill the training was given in small groups (4-8 persons) by two trainers experienced in the area concerned. The content of the training was based on national guidelines for general practice[25,26], with supervised hands-on practice of skills forming the core of each training-session. Total training time was three hours, with one hour for injection technique of the shoulder and ophtalmoscopy, and half an hour for PAP-smear and laboratory examination of fluor. With an objective structured clinical examination[27] proficiency was assessed in the four different clinical skills.

To evaluate the effect of the training a multiple choice test (with 15 items for each skill) was used to measure relevant knowledge. The content of the multiple choice test was developed according to national guidelines. The effect of the training on performance in practice was measured with a log diary, which allowed registration of quantitative and some qualitative criteria of the four different procedures. The criteria were chosen because they were considered sensitive to changes in performance and their registration was feasible.

A small incentive (£75) was provided for completing both registration periods. As an illustration examples of the formats are given in figure 1.

Figure 1 Examples of items on knowledge test and log diary for shoulder injection technique

A. Knowledge test

item 16	Pain in the shoulder with irradiation into the hand is more frequently caused by a problem in the neck rather than the shoulder.	(true/false/don't know)
item 17	Inflammation of the shoulder joint capsule causes disturbances *both* in active and passive examination of shoulder movements.	(true/false/don't know)
item 18	Restriction of the passive horizontal adduction of the arm during physical examination of the shoulder is an indication of osteo-arthritis of the acromio-clavicular joint.	(true/false/don't know)
item 29	A correct insertion point for injection of the glenoid cavity is about 4 cm below the angle between spina scapulae and acromion.	(true/false/don't know)
item 30.	During injection of the glenoid cavity from behind the correct direction of the needle is towards the top of the coracoid process.	(true/false/don't know)

B. Log diary technical skills

Patient code ☐☐☐☐ GP code ☐☐☐☐
Date of Birth ☐☐☐☐☐☐ Date ☐☐☐☐☐☐
☐ male ☐ female

☐ *Injection shoulder*
- ☐ acromio-clavicular joint
- ☐ bursa subacromialis
- ☐ glenohumeral joint
- ☐ other :.....

- ☐ 1st injection
- ☐ repeated injection

Study design

The knowledge test was administered to all participants at the start, serving as a baseline measurement. All participants received personal written feedback on their scores, comparing results with that of their peers, and received an educational handout concerning the four skills, including step-by-step guidelines for procedures. They were subsequently requested to record how often they performed the four different procedures (shoulder injection, ophthalmoscopic control of diabetes mellitus, PAP-smear and laboratory examination of fluor vaginalis) in a log diary covering twenty days. This registration time was considered to be feasible and adequate for detecting group effects in performance. Three months after baseline measurement group A received the training including assessment of skill after the training. Both group A and B were assessed on knowledge. Again all participants received personal feedback on their scores. One month after the course all participants were requested to record performance of procedures in practice covering a second period of twenty days (figure 2).

Figure 2 Design of study

	group A (n = 31)	group B (n = 28)
0	- knowledge test*	- knowledge test*
	educational handout	educational handout
1-2 months	first registration of performance in practice with log diary covering 20 days	
3 months	**short course** - knowledge test*	---- - knowledge test*
4-5 months	second registration of performance in practice with log diary covering 20 days	

* personal feedback was provided on results after each test

Data management and analysis

Scores on the test-formats were converted to percentage of maximum score. T-test was used to compare mean scores between groups, and paired T-test between different periods within groups. Chi-square (or Fisher exact) was used to analyse differences in characteristics and performance between both groups. One-way analysis of variance (with Student-Neuman Keuls as post hoc multiple comparison method) was used to evaluate the influence of nominal and ordinal characteristics on scores.

Results

Only participants who completed log diaries in both registration periods were included in the study. Results for analysis were available from 31 participants of group A and from 28 persons

of group B. On personal and background characteristics group A and group B no statistically significant differences were found (table 1). One participant of group B failed to complete the knowledge test at three months.

Table 1 Personal and practice characteristics*

	group A (n=31)	group B (n=28)	**
Age (SD)	41.3 (5.2)	41.7 (7.3)	ns [a]
Male	24 (80.0)	26 (86.7)	ns
Female	6 (20.0)	4 (13.3)	ns [b]
Years of experience (SD)	11.6 (6.4)	12.5 (8.3)	ns [a]
Practice characteristics			
Solo	17 (56.7)	14 (46.7)	ns
Duo	8 (26.7)	11 (36.7)	ns
Group	5 (16.6)	5 (16.6)	ns [b]
city (>30.000 inh.)	15 (50.0)	9 (30.0)	ns
town (5.000-30.000 inh.)	8 (26.7)	12 (40.0)	ns
rural (<5.000 inh.)	7 (23.3)	9 (30.0)	ns
Number of patients in practice (SD)	2653 (733)	2640 (661)	ns [a]

*figures between brackets are percentages unless stated otherwise
** X^2 test unless stated otherwise ([a] T-test; [b]Fisher exact test)

Table 2 Knowledge test scores* for intervention- and control group (baseline measurement and after three months)

	group A*				group B*				
	n	mean	SD	paired T-test	n	mean	SD	paired T-test	T-test
PAP-smear									
baseline	31	54.6	13.5		28	52.9	11.0		ns
after 3 months	31	68.6	13.0	<0.001	27	61.5	12.5	<0.05	<0.01
Laboratory examination of Fluor vaginalis									
baseline	31	61.1	14.4		28	52.6	12.2		<0.05
after 3 months	31	84.7	11.1	<0.001	27	64.0	16.2	<0.01	<0.001
Shoulder injection									
baseline	31	70.5	16.1		28	71.9	19.2		ns
after 3 months	31	92.9	5.9	<0.001	27	79.8	17.8	<0.05	<0.001
Ophthalmoscopic control in Diabetes mellitus									
baseline	31	43.0	13.9		28	42.1	16.7		ns
after 3 months	31	75.5	13.2	<0.001	27	61.5	24.2	<0.01	<0.01

* All entries expressed as precentage of maximum scores

Competence

Knowledge scores showed no statistically significant differences between both groups at the start, except for laboratory examination of fluor, which was higher in group A (table 2). Both groups showed improvement in scores after three months but improvement was significantly higher for group A (after receiving the training) in all four skills. The performance-based test in group A showed high mean scores for Shoulder injection (85%) and PAP-smear (81%), and more modest scores for Ophtalmoscopic control of diabetes mellitus (73%) and Laboratory examination of fluor vaginalis (70%),

Performance in practice

The performance in practice of the different procedures is shown in table 3. For PAP-smear the number of procedures performed was higher in the second registration period for both groups. The use of recommended materials for taking PAP-smears showed a significant increase in group A. The number of smears without endocervical cells showed no significant differences between periods or groups. The number of examinations of fluor was smaller in the second period in both groups. The proportion of requests for Chlamydia diagnostics did not differ between groups or between periods. Group A performed more shoulder injections after training, while in group B no difference between first and second period was found. Both groups performed few ophtalmoscopic controls in the first and second period.

Table 3 Performance in practice (logdiary entries in 20 days)

		group A*		group B*		
	Period	Sum-score	Difference within A**	Sum-score	Difference within B**	Difference between A-B**
PAP-smear						
procedures	first	349		371		ns
	second	434	<0.01	443	<0.05	ns
use of recommended	first	244 (70.0)		238 (64.3)		ns
collection materials	second	389 (89.7)	<0.01	328 (74.1)	ns	<0.01
smears of poor quality	first	25 (7.2)		23 (6.2)		ns
	second	23 (5.3)	ns	34 (7.7)	ns	ns
Laboratory examination of fluor vaginalis						
procedures	first	93		91		ns
	second	67	<0.05	66	<0.05	ns
requests for chlamydia-	first	32 (34.4)		26 (28.5)		ns
diagnostics	second	22 (32.8)	ns	17 (25.8)	ns	ns
Shoulder injection						
procedures	first	80		90		ns
	second	120	<0.01	72	ns	<0.001
Ophthalmoscopic control in Diabetes mellitus						
procedures	first	29		17		ns
	second	17	ns	8	ns	ns

* Sum-score adjusted to 30 members for both groups to allow comparison (proportions between brackets). ** X^2

Discussion

In this study a significant effect of a skills training on competence for all four topics was demonstrated. An effect of the training on performance in practice was found only for shoulder injection and PAP-smear, whereas no such effect could be demonstrated for opthalmoscopic control and laboratory examination of fluor.

These results support the notion that a good degree of competence is a necessary but not always a sufficient condition for changing physician performance in the practice setting.

For shoulder injection the practical skills training was effective to facilitate use of the procedure in practice in group A, while in group B, although all the necessary information was in the handout, without practical training no increased application of the procedure was observed. For PAP-smear no difference in the number of procedures was found between groups. This is perhaps not surprising, given that the screening program for cervical cancer in the Netherlands was, at the time of the study, organised by public health authorities and not dependent on the initiative of the general practitioner. However, there was substantial improvement in the use of recommended material for collection of PAP-smear after training. It is remarkable that demonstration of and practising with recommended material was sufficient to generate such a change in the use of materials, without having additional support. No statistically significant improvement of quality of PAP-smears could be demonstrated (although there was a clear tendency), as the quality was already very high and numbers were too small. The training in examination of fluor vaginalis, also highlighting clues to request chlamydia diagnostics, failed to result in increased use of the procedure. Training time may have been too short. Moreover, the procedure requires considerable time - which is scarce during surgery hours - and adequate organisation of equipment and materials. So organisational obstacles may hinder application of this procedure. Training in opthalmoscopic control of diabetes mellitus also failed to result in increased application in daily practice. Others have indeed demonstrated that opthalmology skills require extensive training[9].

Some methodological aspects in this study have to be taken into consideration for the interpretation of the results. First, the validity of self-reported performance may be questioned, due to over- or underreporting. We did not systematically check log diary entries but all registrations required retraceable patient codes, and on occasional checks no indications were found of reporting a procedure which had not been performed. There is some indication of underreporting in the second period as the number of performances decreased in both the experimental and control group for fluor and opthalmoscopy, so small changes may have been missed. Further, it is clear that quantity of performance as sole endpoint does not necessarily give an indication of quality of performance in practice. Conclusions on quality of performance must therefore be drawn with some caution, although skills test results were reassuring after training. Hypothetically it is possible that quality in performance improved while quantity decreased and vice versa. The selection of quantitative and some qualitative endpoints to evaluate the effect of the course was for logistical and financial reasons. Direct observation of procedures in practice or more detailed information on outcome would be preferable, but

require substantial resources.

In conclusion, while for some skills a short focussed training seems adequate to bring about desired changes, for other skills more complex interventions are needed. CME needs to consider which interventions work and which don't for each topic seperately.

References

1. Spike N, Veitch C. Procedural skills for general practice. Aust Fam Physician 1990;19:1545-53.
2. Heikes LG, Gjerde CL. Office procedural skills in family medicine. J Med Educ 1985;60:444-53.
3. Lamberts H, Brouwer H, Mohrs J. Reason for encounter- episode- and process-oriented standard output from the transition project. Amsterdam: Department of General Practice, University of Amsterdam, 1991.
4. Tan LHC. Tekorten in de opleiding van huisartsen (Deficiencies in vocational training of general practitioners). Dissertation (with English summary). Amsterdam: University of Amsterdam, 1989.
5. Spike N, Veitch C. Competency of medical students in general practice procedural skills. Aust Fam Physician 1991;20:586-91.
6. Patrick J. Training: Research and Practice. London: Academic Press, 1992.
7. Campbell HS, Fletcher SW, Lin S, Pilgrim CA, Morgan TM. Improving physicians' and nurses clinical breast examination: a randomized controlled trial. Am J Prev Med 1991;7:1-8.
8. Berden HJJM. Basic Cardiopulmonary Resuscitation. Assessment of skills in training situations. Dissertation. Utrecht: University of Utrecht, 1993.
9. Reenders K, De Nobel E, Van den Hoogen HJM, van Weel C. Screening for diabetic retinopathy by general practitioners. Scand J Primary Health Care 1992;10: 306-9.
10. Jansen JJM, Scherpbier AJJA, Metz JCM, Grol RPTM, Van der Vleuten CPM, Rethans JJ. Construct validity of performance-based assessment in continuing medical education for general practitioners. Med Educ 1996;30: 339-44.
11. Fisher EW, Pfleiderer AG. Assessment of otoscopic skills of general practitioners and medical students: is there room for improvement? Br J Gen Pract 1992;42:65-7.
12. Roland MO, Porter RW, Matthews JG, Redden JF, Simonds GW, Bewley B. Improving care: a study of orthopaedic outpatient referrals. Br Med J 1991;302:1124-8.
13. Haynes RB, Davis DA, McKibbon A, Tugwell P. A critical appraisal of the efficacy of continuing medical education. JAMA 1984;251:61-4.
14. McLaughlin PJ, Donaldson JF. Evaluation of continuing medical education programs: a selected literature, 1984-1988. J Continuing Educ Health Professions 1991;11:65-84.
15. Davis DA, Thomson MA, Oxman AD, Haynes B. Evidence for the effectiveness of CME. A review of 50 randomized controlled trials. JAMA 1992;268:1111-7.
16. Tamblyn R, Battista R. Changing clinical practice: which interventions work? J Continuing Educ Health Professions 1993;3:273-88.
17. Wensing M, Grol R. Single and combined strategies for implementing changes in primary care: a literature review. Quality in Health Care 1994;6:115-132.
18. Grol R. Implementing guidelines and changes in practice. Quality in Health Care 1992;1:184-191.
19. Robertson N, Baker R, Hearnshaw H. Changing the clinical behaviour of doctors: a psychological framework.

Quality in Health Care 1996;5:51-4.

20. Kanouse DE, Kallich JD, Kahan JP. Dissemination of effectiveness and outcomes research. Health Policy 1995; 34:167-92.

21. Stein LS. The effectiveness of continuing medical education: eight research reports. J Med Educ 1981;56:103-10.

22. Vierhout WPM, Knottnerus JA, Van Ooij A, Crebolder HFJM, Pop P, Wesselingh-Megens AMK, et al. Effectiveness of joint consultation sessions of general practitioners and orthopaedic surgeons for locomotor-system disorders. Lancet 1995;346:990-4.

23. Dekker JH, Boeke JHP. Vaginale klachten in de huisartspraktijk [Vaginal complaints in general practice] Dissertation (with English summary). Amsterdam: Vrije Universiteit, 1992.

24. Boon ME, Alons-van Kordelaar JJM, Rietveld-Scheffers PEM. Consequences of the introduction of combined spatula and cytobrush sampling for cervical cytology. Improvements in smear quality and detection rates. Acta cytologica 1986;30:264-9.

25. Grol R. National standard setting for quality of care in general practice: attitudes of general practitioners and response to a set of standards. Br J Gen Pr 1990;40:361-4.

26. Rutten GEHM, Thomas S (eds). NHG-standaarden voor de huisarts [national guidelines for the general practitioner]. Utrecht: Bunge, 1993.

27. Harden RM, Gleeson F. Assessment of clinical competence using an objective structured clinical examinations. Med Educ 1979;13:41-54.

Hoofdstuk 8

Failure of feedback to enhance self-assessment skills of general practitioners*

Summary

Background. Self-directed learning requires accurate self-assessment, but research evidence shows poor validity of self-assessment. Training in self-assessment may improve validity.
Purpose. To investigate if repeated personal feedback based on objective knowledge and skill scores enhances self-assessment skills of practicing general practitioners.
Method. Subjects were general practitioners (n=60), who received a skills training covering four clinical skills at three months (group A) or six months (group B) after enrollment in the study. Participants were tested at three month intervals with a knowledge-test (60 items), a performance based test (4 stations) and a self-assessment questionnaire (22 items), covering the four different clinical skills. They received personal feedback on the results.
Results. At three months mean scores on the self-assessment questionnaire and knowledge test had increased significantly more in group A compared to group B, while at six months no differences in mean scores remained. Correlations between self-assessment rating and objective scores were low to moderate, with little overall improvement over time.
Conclusion. While self-assessment scores can to some extent be useful in measuring perceived changes in competence in groups, individual self-assessment scores on its own are an invalid source of information concerning competence of practicing physicians, and that this does not improve significantly with regular feedback.

Introduction

Self-directed learning is the dominant mode of learning for professionals after graduating. It implies a process in which individuals take the responsibility for diagnosing their learning needs. Accurate self-assessment, the ability of physicians to perceive areas of strength and weakness in their competence or performance, is therefore considered an essential requisite for effective adult learning,[1,2] and the development and maintenance of professional competence.[3,4] While the importance of self-assessment is widely recognized, research evidence has provided little support for validity of self-assessment in relation to expert ratings or objective tests.[5,6] Correlations between self-assessment and expert ratings among students at different levels of expertise and for various aspects of clinical competence are generally low

---

* Accepted for publication as: Jansen JJM, Van der Vleuten CPM, Grol RPTM, Crebolder HFJM, Rethans JJ. Failure of feedback to enhance self-assessment skills of general practitioners

to moderate.[7-12] It has been argued that these results are a consequence of the absence of specific training in self-assessment skills in medical training programs, and that validity and accuracy of self-assessment may improve with self-assessment training.[5,13]

Personal feedback on actual achievement has proven to be a powerful method of training,[14] but few studies have investigated development of self-assessment skills over time with subjects receiving regular feedback on actual scores. Cochran and Spears[15] and Hay[16] found correlations between student and instructor-rating increase from moderate to high. Regular feedback opportunities, including discussion of the completed self-evaluations, were part of the course. However, results were not compared to objective tests, and the high correlations can be explained by either improved self-evaluation skills or, alternatively, as evidence of successful negotiation between students and instructors.[15,16]

Only two studies, both among undergraduate medical students, have investigated the development of self-assessment over time, comparing self-assessment with objective measures. Arnold et al. compared self-ratings of medical students during four years with faculty-ratings.[17] Both student- and faculty-rating showed annual increase, with students' increase being smaller. Correlation between self- and faculty-rating decreased to a nonsignificant level among more advanced students, while no correlation was found between self-assessment rating and objective test scores. Students with higher scores on objective tests were more conservative in their self-assessments. Rezler studied the development of self-assessment among medical students during their first two years in a problem-based curriculum.[18] Mean ratings increased from first to second year, but correlations dropped to nonsignificant levels. Student self-ratings for reasoning and knowledge showed no significant correlation with a knowledge test administered in the second year.

Few studies have investigated self-assessment skills of experienced physicians. Because physicians frequently base their decisions to attend continuing medical education on self-assessment it is important to know if they can learn to become more accurate in assessing their own educational needs. The present study investigated whether repeated personal feedback based on objective knowledge- and skill-scores for technical procedures performed in primary care, enhanced self-assessment skills of practicing general practitioners with regard to these procedures.

Method

General practitioners were invited to participate in a continuing medical education (CME) course on technical clinical skills as part of an experiment investigating transfer of skills from training to practice environment. Those who agreed to participate, were divided into two groups, according to their preference for the moment of the course. One group (A) received the course three months after enrollment, whereas the other group (B) received the course after

six months.

The course covered four different skills: injection technique of the shoulder, ophthalmoscopic control in diabetes, PAP-smear and laboratory examination of fluor vaginalis. The topics were identified by general practitioners as having priority and selected by course providers because proficiency in these procedures was known to be amenable to improvement. The objective of the course was to increase relevant knowledge and proficiency in performance of procedures according to national guidelines for general practice. For each skill the training was given in small groups (4-8 persons) by two trainers experienced in the area. The content of the training was based on national guidelines for general practice, and included discussion of the guidelines with supervised hands-on practice of skills forming the core of each training-session. Total training time was three hours, with one hour for injection technique of the shoulder and ophtalmoscopy, and half an hour for PAP-smear and laboratory examination of fluor. Satisfaction of the participants was measured directly after the course using a questionnaire.

To evaluate the cognitive effect of the training a 60-item multiple choice test was used to measure relevant knowledge. The content of the knowledge test was directly based on the content of the course and using national consensus guidelines[19]. With an Objective Structured Clinical Examination using trained observers and detailed checklists, with 23-33 items for each skill based on the content of the course, proficiency was assessed in the four different clinical skills. Manikins were used for shoulder injection (Limbs & Things® shoulder model) and PAP-smear (Schultz ®). For fluor vaginalis a specimen of fluor was used, while trained real patients were used for opthalmoscopic control in Diabetes Mellitus. A self-assessment questionnaire was developed, based on the content of the course, and consisted of 22 items to be scored on a seven-point Likert scale. As an illustration, sample items of the different formats are shown in figure 1.

The knowledge test and self-assessment questionnaire were administered at the start, after three months and after six months. The performance based test was administered after three months (intervention-group only) and six months. All participants received personal detailed written feedback on their scores. The correct answers were provided, allowing review of errors, and individual scores for each procedure were compared with results of the peers, indicating whether scores were low, average or high. With the feedback written educational information was provided reviewing correct performance, with step-by-step guidelines for the different procedures.

Scores on the test-formats for the different skills were aggregated to total scores and converted to a percentage of the maximum score. To evaluate self-assessment in relation to objective scores the differences between self-assessment scores and knowledge or skill scores were calculated after transforming original scores into Z-scores to adjust for differences in average scores. Subgroups of low, intermediate and high scores were constructed by equally dividing participants over the three subgroups with group assignment according percentile score. T-test

Figure 1 Examples of items on self-assessment questionnaire, knowledge test and performance-based test for shoulder injection technique

Self assessment questionnaire

		Anatomy 1 = unable to demonstrate; 2 = very elementary, unreliable interpretation; 3 = some impression, insufficient for practical use; 4 sufficient competence for practical use; 5 sufficient confidence for practical use, some errors; 6 good knowledge and routine; 7 excellent knowledge							*educational need* 1 = not; 2 = small; 3 = may be of some interest; 4 = strongly; 5 = considerable, more than average; 6 = large; 7 = very large						
1.	anatomy of the shoulder joint	1	2	3	4	5	6	7		1	2	3	4	5	6	7
2.	Physical Ex.of the shoulder joint	1	2	3	4	5	6	7		1	2	3	4	5	6	7
3.	Differential Diagnosis of shoulder complaints	1	2	3	4	5	6	7		1	2	3	4	5	6	7
4.	injection technique articulatio glenohumeralis	1	2	3	4	5	6	7		1	2	3	4	5	6	7
5.	injection technique bursa subacromialis	1	2	3	4	5	6	7		1	2	3	4	5	6	7
6.	injection technique subacromial space	1	2	3	4	5	6	7		1	2	3	4	5	6	7
7.	injection technique articulatio acromio-clavicularis	1	2	3	4	5	6	7		1	2	3	4	5	6	7

Knowledge test

item 16 Pain in the shoulder with irradiation into the hand is more frequently caused by a problem in the neck rather than the shoulder. (true)

item 17 Inflammation of the shoulder joint capsule causes disturbances *both* in active and passive examination of shoulder movements. (true)

item 18 Restriction of the passive horizontal adduction of the arm during physical examination of the shoulder is an indication of osteoarthritis of the acromio-clavicular joint. (true)

item 29 A correct insertion point for injection of the glenoid cavity is about 4 cm below the angle between spina scapulae and acromion. (false)

item 30 During injection of the glenoid cavity from behind the correct direction of the needle is towards the top of the coracoid process. (true)

Performance-based test (part of scoring grid)

	not/incorrectly performed	correct performed
Choice of materials and preparation for injection		
- correct syringe	☐	☐
- adequate needle	☐	☐
- adequate dosage of lignocaine	☐	☐
- adequate dosage of corticosteroid	☐	☐
- disinfection of skin and palpating fingers	☐	☐
Injection technique		
- correct insertion site	☐	☐
- correct angle of insertion	☐	☐
- correct depth	☐	☐
- aspiration before injection	☐	☐

was used to compare mean scores between intervention- and control group, and paired T-test between different moments. Chi-square (or Fisher exact) was used to analyze differences in percentages between groups. One-way analysis of variance (with Student-Neuman Keuls as post hoc multiple comparison method) was used to evaluate the influence of nominal and ordinal characteristics on scores. For interrater reliability of the performance-based test intra-class coefficients were used.[20] Bivariate correlations were expressed as Pearson product moment coefficients.

Results

Scores were available for 60 participants, equally divided between group A and B. However, from various participants no complete data were available. At the start one knowledge test scoring sheet from group B was not returned. At six months 24 participants of group A were tested on skill and only 19 knowledge test scores were available, while all but one filled out the self-assessment rating sheet. Subgroup analysis of non-participants versus participants at six months revealed no indication of selection bias. Personal characteristics (age, sex, years of experience) and practice characteristics of both groups showed no statistically significant differences. Satisfaction of participants was high about course content (97%), assessment (89%) and feedback (93%) and not significantly different between both groups.

Table 1 Knowledge-, skill- and self-assessment scores for group A and B at three-month intervals

	knowledge			skill			self-assessment		
	n	mean	SD	n	mean	SD	n	mean	SD
begin-score									
group A	30	51.5	7.9	not administered			30	50.0	10.8
group B	29	49.0	6.7	not administered			30	50.6	10.4
after 3 months									
group A	30	80.1	6.4 [a,b]	30	75.7	9.3	30	65.9	9.9 [a,b]
group B	30	66.3	11.2 [b]	not administered			30	53.2	13.4
after 6 months									
group A	19	81.5	9.0	24	81.8	5.6	29	70.3	10.5
group B	30	79.8	9.5 [b]	30	80.0	7.2	30	67.6	10.4 [b]

* All entries expressed as percentage of maximum score [a] $p < 0.001$ between group A and B [b] $p < 0.001$ between begin-score vs 3 months or 3 months vs 6 months within groups

Scores on knowledge and self-assessment showed no statistically significant differences between groups at the start (table 1). Both groups showed improvement in scores after three months but improvement was significantly higher in group A. At six months, after group B had also received the training, no significant differences in scores remained between the two groups.

No significant influences of personal and practice characteristics on knowledge- skill- and self-assessment score were found, except for city practices, whose doctors had higher self-assessment scores (but not higher knowledge- or skill scores) at three and six months.
The interrater reliability for the performance-based test was 0.80 at three months and 0.83 at six months. Correlations between self-assessment rating and knowledge-score were low at the start, increased to moderate at three months, and declined again at six months to the same level as at the start. The correlation between self-assessment and skill was very low at three months and increased to moderate levels at six months (table 2).

Table 2 Correlations between knowledge- skill- and self-assessment scores

	N	Knowledge	Self-assessment
at start			
Knowledge	59		0.19
after three months			
Knowledge	60		0.46 [a]
Skill	30	0.24	0.06
after six months			
Knowledge	49		0.21
Skill	54	0.20	0.46 [a]

[a] $p<0.001$

Accuracy of self-assessment was additionally assessed by comparing the mean standardized difference between self-assessment and knowledge- or skill score, after equally dividing participants in three subgroups (low, intermediate and high) using the total scores on the knowledge test and skills test. The results are given in table 3. The mean standardized difference scores between self-assessment and knowledge or skill score of the low scoring group varied between 0.33 and 0.74 above the overall mean, while the mean difference for the high scoring group ranged between 0.30 and 0.72 below the overall mean. Variance within groups (expressed as standard deviations) was large compared to differences between groups.

Table 3 Difference between self-assessment score and scores on the knowledge and skills test* for low, medium and high achievers**

		low achievers		medium achievers		high achievers		
Variable	N	Mean	SD	Mean	SD	Mean	SD	F probability
at start								
Knowledge	59	+0.72	0.88	- 0.16	0.84	- 0.53	0.85	<0.001
at three months								
Knowledge	60	+0.43	1.22	- 0.12	0.86	- 0.31	0.75	<0.05
Skill	30	+0.74	1.00	- 0.02	0.56	- 0.72	0.85	<0.01
at six months								
Knowledge	49	+0.68	0.78	+0.26	0.67	- 0.84	0.86	<0.001
Skill	54	+0.33	1.06	- 0.04	1.02	- 0.30	0.85	ns

* All scores after Z-transformation of difference between self-assessment score and knowledge test or skills test score. ** Group assignment according percentiles: low achievers: scores < p 33.3; medium achievers: scores p33.3-p66.7; high achievers: scores > p 66.7.

Discussion

The increase in mean self-assessment scores among the participants, and the different patterns of increase of the intervention-group compared to the control-group, together with corresponding changes in knowledge-score, can be considered supportive for validity of self-assessment at the group level. Growth in self-evaluation scores during training have been reported in various studies.[16-18, 21,22,24] However, in isolation, this is hardly compelling evidence for the validity of self-evaluation.

The results of correlation-analysis suggest some improvement of self-assessment over time, first for knowlegde and later for skill. This can be interpreted as an indication that participants shifted from predominantly 'knows how' (at three months) to 'shows how' (at six months),[25] and thus became somewhat more accurate in their self-assessment of proficiency concerning the technical clinical skills. Nevertheless, self-assessment was generally a poor predictor of competence.[26] Only 20-25% of variance on self-assessment was explained by the scores on the objective tests, and no substantial effect of repeated detailed personal feedback could be demonstrated.

Other factors apparently heavily influenced the relation between self-assessment and competence. We found no indications of significant contributions of personal characteristics (age or sex), professional (experience) or practice characteristics, with exception of rural-urban

differences at three and six months. These results are consistent with other research.[17,27,28] Some authors have argued that self-assessment is strongly influenced by noncognitive attributes, such as self-representation and personality.[6,7,17,23,29] These factors might account for the large unexplained variability of self-assessment in this study.

The finding that high achievers tended to underestimate their knowledge or performance, while low achievers tended to overestimate, is consistent with the results from other self-assessment studies.[5,7,10,15,23,30-32] Various explanations have been forwarded for this intriguing phenomenon. One possible explanation is that the physicians were evaluating themselves according their ambition rather than their actual performance,[7] implicating that high achievers may compare themselves to more stringent standards.[17] Alternatively they may have viewed their performance as not being quite as good as it was influenced by their internal self-representations developed early in life.[23] However, the consistent finding could also be interpreted as yet another proof of poor validity of self-assessment. As the statistical chance for low-achievers to overestimate their knowledge and performance is greater than for high achievers, these results may in fact underline that self-assessment is of questionable validity.

This study has some limitations. The participants are a rather small group of physicians who volunteered to submit themselves to an experiment including assessment. Although we found no indication that non-participants at six months differed from their peers, the missing of knowledge scores from 11 participants and skill scores from 6 participants in the intervention group at six months are a potential source of bias in this study. Another limitation was that participants were not very familiar with self-assessment, so scores may reflect a large error component. Nevertheless it would have been likely to expect that some learning would occur during the course, especially in this highly motivated group, resulting in improved correlations between self-assessment scores and objective scores, and this effect was not observed. The use of pre- and postintervention self-assessments can be criticised because internal standards of participants may change as a result of the intervention, and instead the use of retrospective pre-post self-assessments has been recommended as more consistent with objective measures.[33,34] However, the skills covered in this study were all familiar to participants. Therefore a misunderstanding or misconcepttion of the skill or concept is not likely. Moreover the use of retrospective pre-post self-assessment would not have changed the self-assessment score at six months, after having received the training, and this score also correlated rather poorly with objective measures. It might be argued that feedback procedures as used in this study were not adequate to realize changes[13]. However such feedback procedures have been proven to be effective to bring about change in competence and performance,[35] and self-assessment scores at group-level did show changes reflecting changes in objective scores, but at the individual level self-assessment failed to become more accurate. So we believe that the findings in this study indeed support the minor influence of feedback on self-assessment, consistent with research indicating that self-assessment is more closely related to generalized self-attributions and only minimally influenced by external feedback.[36]

This result has important implications for a voluntary continuing medical education system in which selection of activities is based on individual preferences. While it is obvious that motivation is crucial in adult learning[37], the use of objective measurement must provide the basis to bring subjective and objective learning needs closer to each other, to enhance rational choice of continuing medical education topics.[38,39,40]

In conclusion, while self-assessment scores at the group level can be useful to some extent in measuring perceived changes in competence, individual self-assessment scores on their own are an invalid source of information concerning competence of practicing physicians, and this does not improve significantly with regular feedback. Therefore objective tests should have a much larger place as a basis for individualized continuing medical education.

References

1. Knowles MS. The modern practice of adult education: from pedagogy to andragogy. Chicago: Follett, 1980.
2. Jennett P, Jones D, Mast T, Egan K, Hotvedt M. The characteristics of self-directed learning. In: Davis DA, Fox RD (eds). The Physician as a learner: linking research to practice. Chicago: American Medical Association, 1994:47-65.
3. Schön DA. The reflective practitioner: how professionals think in action. New York: Basic Books, 1983.
4. Schön DA. Educating the reflective practitioner: towards a new design for teaching and learning in the professions. San Francisco: Jossey-Bass, 1987.
5. Boud D, Falchikov N. Quantitative studies of student-self-assessment in higher education: a critical analysis of findings. Higher Educ. 1989;18:529-49.
6. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 1991;66:762-9.
7. Stuart MR, Goldstein HS, Snope FC. Self-evaluation by residents in family medicine. J. Fam Pract 1980;10:639-42.
8. Scalabassi SE, Woelfel SK. Development of self-assessment skills in medical students. Med Educ 1984;18:226-31.
9. Kolm P, Verhulst S. Comparing self- and supervisor evaluations. A different view. Eval Health Professions 1987;10:80-9.
10. Cushing AM, Jolly BC, Dacre JE, Hitman G, Griffiths S, Southgate L. Critical self-appraisal and examiner, patient and student self-ratings of communication skills in OSCE. In: Rothman AI, Cohen R (eds). The sixth Ottawa conference on medical education. Toronto: University of Toronto, 1994:145-8
11. Furman GE, Colliver JA, Galofré A. Student self-ratings and standardized patient ratings of a medical interview. In: Rothman AI, Cohen R (eds). The sixth Ottawa conference on medical education. Toronto: University of Toronto, 1994:142-4.
12. Jansen JJM, Tan LHC, Van der Vleuten CPM, Van Luijk SJ, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners. Med Educ 1995;29:247-53.
13. Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med 1992;67:672-9.
14. Patrick J. Training: Research and Practice. London: Academic Press, 1992.

15. Cochran SB, Spears MC. Student self-assessment and instructors' ratings: a comparison. J Am Diet Assoc 1980;76:253-5.

16. Hay JA. Investigating the development of self-evaluation skills in a problem-based tutorial course. Acad Med 1995;70:733-5.

17. Arnold L, Willoughby TL, Calkins EV. Self evaluation in undergraduate medical education: the longitudinal perspective. J Med Educ 1985;60:21-8.

18. Rezler AG. Self-assessment in problem-based groups. Med Teach 1989;11:151-6.

19. Grol R. National standard setting for quality of care in general practice: attitudes of general practitioners and responses to a set of standards. Br J Gen Pract 1990;40:361-4.

20. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981;20:111-23.

21. Bleys FC, Gerritsma JGM, Netjes I. Skills development by medical students and the influence of prior experience: a study using evaluation by students and self-assessment. Med Educ 1986;20:234-9.

22. Woolliscroft JO, Palchik NS, Dielman TE., Stross JK. Self-evaluation by house officers in a primary care training program. J Med Educ 1985;60:840-6.

23. Woolliscroft JO, Tenhaken J, Smith J, Calhoun JG. Medical students' clinical self-assessments: comparisons with external measures of performance and the students' self-assessments of overall performance and effort. *Acad Med* 1993;68:285-94.

24. Day SC, Cook EF, Nesson HR, Wolf MA, Goldman L. A learning-curve approach to the self-assessment of internal medicine training. J Med Educ 1984;59:672-5.

25. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1993;65:S63-7.

26. Norman GR. Defining competence: a methodological review. In: Neufeld VR, Norman GR. Assessing Clinical Competence. New York, Springer, 1985:15-35.

27. Strober Escovitz E, Cohen DG. An analysis of gender differences in students clinical competence self-assessments. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra R P (eds). Teaching and assessing clinical competence. Groningen: Boekwerk, 1990:110-7.

28. Farmer E. Changes in self assessment of communication skills of Australian family medicine programme trainees compared with a group of hospital peers. In: Harden RM, Hart IR, Mulholland H (eds). Approaches to the assessment of clinical competence. Dundee: Centre for Medical Education, 1992:121-5.

29. Kegel-Flom P. Predicting supervisor, peer and self-ratings of intern performance. J Med Educ 1975;50:812-5.

30. Richards BF, Philip EB, Frye AW, Philip JR. Integrating self-assessment in an OSCE for pre-clinical medical students. In: Hart IR, Harden RM, Des Marchais J (eds). Current developments in assessing clinical competence. Montreal: Can-Heal, 1992:338-46.

31. Nathan RG. Using an in-training examination to assess and promote the self-evaluation skills of residents. Acad Med 1992;67:613.

32. Morton JB, Macbeth WAAG. Correlations between staff, peer and self assessments of fourth-year students in surgery. Med Educ 1977;11:167-70.

33. Levinson W, Gordon G, Skeff K. Retrospective versus actual pre-course self-assessments. Eval Health Professions 1990;13:445-52.

34. Skeff KM, Stratos GA, Bergen MR. Evaluation of a medical faculty development program: the comparison of traditional pre/post an retrospective pre/post self-assessment ratings. Eval Health Professions 1992;15:350-66.

35. Wensing M, Grol R. Single and combined strategies for implementing changes in primary care: a literature review. Quality in Health Care 1994;6:115-32.

36. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 1991;66:762-9.

37. Mann K, Ribble J. The role of motivation in self-directed learning. In: Davis DA, Fox RD (eds). The Physician as a learner: linking research to practice. Chicago: American Medical Association, 1994: 69-88.

38. Sibley JC, Sackett DL, Neufeld V, Gerard B, Rudnick KV, Fraser W. A randomized trial of continuing medical education. N Engl J Med 1982; 306: 511-5.

39. Premi J. Individualized continuing medical education. In: Davis DA, Fox RD (eds). The Physician as a learner: linking research to practice. Chicago: American Medical Association, 1994: 201-16.

40. Tracey JM, Arroll B, Richmond DE, Braham PM. The validity of general practitioners' self-assessment of knowledge: cross sectional study. Br Med J 1997;315:1426-8.

Hoofdstuk 9

Waardering en kosten van vaardigheidstraining en toetsing in de huisartsgeneeskunde

Inleiding

Voor toepassing van toetsing zijn naast validiteit en betrouwbaarheid ook de acceptatie en haalbaarheid van toetsingsmethoden van groot belang (Hays 1994). De acceptatie van vaardigheidstoetsing in het medisch curriculum blijkt goed te zijn (Newble 1988; Driessen 1987). Over de acceptatie van vaardigheidstoetsing bij nederlandse practizerende artsen zijn geen gegevens bekend. Uit ervaringen met vaardigheidstoetsing in Nederland (Tan 1988; Pollemans en Tan 1990; Jansen et al. 1995; Jansen et al. 1996) en internationaal (Reznick et al 1992; Hays et al. 1993; Grand'Maison et al. 1992) is gebleken dat deze vorm van toetsing organisatorisch relatief complex is, maar in allerlei varianten inmiddels met succes is beproefd. De kosten blijken sterk te variëren, afhankelijk van de plaatselijke omstandigheden, en bedragen doorgaans NLG 65-100,- per kandidaat per vaardigheid (Reznick et al. 1993; Cusimano et al 1994; Carpenter 1995). Voor de Nederlandse situatie zijn geen gegevens over kosten bekend. In dit hoofdstuk wordt ingegaan op de waardering en kosten van dergelijke nascholing.
De volgende vraagstellingen werden geformuleerd:
- Hoe is de acceptatie onder huisartsen van educatieve toetsing van technische vaardigheden?
- Welke vorm van toetsing heeft de meeste voorkeur?
- Wat zijn de kosten voor toetsing van vaardigheden?

Methode

De waardering voor vaardigheidstraining en toetsing werd gemeten in een drietal experimentele situaties. In het eerste experiment vond uitsluitend toetsing plaats van acht verschillende technische vaardigheden in een vaardigheidstoets (fundoscopie, catheterisatie, mictieklachten, pijn op de borst, pijnlijke enkel, plaatsing IUD, reanimatie, verminderd gehoor). Tevens werd een kennis-over-vaardighedentoets en zelf-beoordelingslijst afgenomen. Deelnemers waren 47 huisartsen-in-opleiding en 49 huisarts-opleiders.
In het tweede experiment werd een vaardigheidscursus gegeven over vier verschillende onderwerpen (reanimatie, onderzoek schouder, injectie schouder en infuusbehandeling). Toetsing vormde onderdeel van de cursus en bestond uit een vaardigheidstoets, een kennistoets over vaardigheden en zelf-beoordeling. Deelnemers waren 72 huisartsen.
Het derde experiment betrof opnieuw een vaardigheidscursus met toetsing als integraal onderdeel over vier verschillende onderwerpen (fundoscopie bij diabetes mellitus type II, injectie schouder, cervix uitstrijk, fluor/SOA-diagnostiek). Toetsing bestond uit een vaardigheidstoets, een kennistoets over vaardigheden en zelf-beoordeling. Deelnemers waren 64 huisartsen.

De waardering van training en toetsing werd gemeten middels een schriftelijke enquete. Deze werd voorgelegd aan de deelnemers van de bovengenoemde drie experimenten. De enquete bestond uit een gedeelte met algemene vragen over de cursus, waarin gevraagd werd naar organisatorische en inhoudelijke aspecten van cursus en/of toets, en de bruikbaarheid van dergelijke toetsing voor de deelnemers. In een tweede gedeelte van de lijst werden vragen gesteld over training van de afzonderlijke onderwerpen die tijdens de verschillende cursussen aan bod kwamen. Voor beantwoording kon een keuze gemaakt worden uit opties op een vijf punts Likert schaal (geheel eens-eens-neutraal-oneens-geheel oneens). De enquête werd anoniem ingevuld direct na afloop van elke cursus.

De organisatie en uitvoeringskosten van de drie cursussen werden berekend op basis van de financieel administratieve gegevens die voor elke cursus werden bijgehouden. De personele kosten van materiaalontwikkeling werden niet in de berekening meegenomen. Alle uitgaven ten behoeve van een cursus werden afzonderlijk geadministreerd. Middels vergelijking met de

Tabel 1 Evaluatiegegevens deelnemers wat betreft algemene aspecten van toetsing. Percentage deelnemers dat aangaf het (geheel) eens te zijn met de voorgelegde uitspraak

	exp 1		exp 2	exp 3
item	haio's n=47	huisartsen n=49	huisartsen n=71	huisartsen n=58
De gang van zaken betreffende de cursus was mij tevoren voldoende duidelijk	63	59	82	93
De organisatie van de cursus verliep soepel	92	90	95	96
Ik vond het prettig om aan de toets/cursus deel te nemen	83	86	96	97
Ik vond de vaardigheden die in de toets/cursus waren opgenomen huisartsgeneeskundig relevant	88	90	97	91
De toetsing vormde een nuttig onderdeel van de cursus	nvt	nvt	76	89
Ik vond het bezwaarlijk bij het verrichten van de vaardigheid te worden geobserveerd	17	8	4	4
Ik heb tijdens de toets goed blijk kunnen geven van mijn vaardigheidsniveau	52	72	74	64
Ik vond het prettig om direct feedback te krijgen van de observator	nvt	nvt	98	93
Ik ervaar de vaardigheidstoets als een nuttige vorm van feedback	62	74	89	97
Ik vond de kennis-over-vaardigheidstoets huisartsgeneeskundig relevant	88	61	82	80
Ik ervaar de kennistoets als een nuttige vorm van feedback	58	61	79	89

posten op de financiële eindverantwoording van het project werd de eigen administratie gecontroleerd op volledigheid. Voor enkele posten, die niet waren doorberekend maar wel een reëel onderdeel vormden van de kosten, werd een inschatting gemaakt.

Resultaten

Van experiment 1 waren evaluatiegegevens beschikbaar van 96 deelnemers. Voor experiment 2 en 3 waren dat respectievelijk 71 en 58 deelnemers.
In tabel 1 en 2 worden de resultaten gepresenteerd wat betreft waardering van toetsing en training, waarbij de respons in de categoriën 'eens' en 'geheel eens' werden samengevoegd. In tabel 1 zijn de algemene vragen over de diverse cursussen opgenomen, terwijl in tabel 2 de waardering voor afzonderlijke cursusonderdelen is vermeld.

Tabel 2: Evaluatiegegevens deelnemers wat betreft afzonderlijke cursus onderwerpen
Percentage deelnemers dat aangaf het (geheel) eens te zijn met de voorgelegde uitspraak

	De inhoud van de training was huisartsgeneeskundig relevant	Er was voldoende tijd om te oefenen	De training had voldoende diepgang
Experiment 2			
Schouderonderzoek en -injectie	97	77	64
Reanimatie en infuustoepassingen	97	94	87
Experiment 3			
Fundoscopie bij diabetes mellitus	79	33	55
Schouderinjectie	97	90	85
Cervix uitstrijk	91	96	92
Fluor/SOA diagnostiek	97	39	73

De deelnemers aan de verschillende experimenten waardeerden de toetsing positief, waarbij de huisartsen in experiment 1 een positiever oordeel hadden over de toetsing dan de huisartsen-in-opleiding. De vaardigheidstoetsing kreeg een hogere waardering dan de kennis-over-vaardigheden-toets. De waardering van de huisartsen was hoger bij experimenten 2 en 3, waar de toetsing deel uitmaakte van de training.
Uit de respons over de verschillende onderwerpen die deel uitmaakten van de cursussen blijkt dat van de meeste onderwerpen de relevantie hoog wordt geacht, met uitzondering van fundoscopie bij diabetes mellitus, welke door relatief veel huisartsen minder relevant wordt gevonden. Voor schouderonderzoek en -injectie in experiment 2, en voor fundoscopie en fluor/soa-diagnostiek in experiment 3 vond een relatief grote groep deelnemers de trainingstijd

en de diepgang tekort schieten.

In tabel 3 is een overzicht opgenomen van de kosten van het organiseren van de diverse experimenten. In experiment 1 werd alleen een toets georganiseerd. Verschillende kostenposten waren daarbij niet expliciet meegenomen in de berekening, zoals de kosten van observatoren en de kosten van drukwerk, zodat daarvan een schatting is gemaakt. De kosten van de observatoren voor experiment 1 werden geschat op NLG 7.500,- op basis van een vergoeding van NLG 150,- per observator per dag, omdat een dergelijke vergoeding ook bij de volgende experimenten aan de observatoren was gegeven. De organisatiekosten (personele kosten) werden geschat op NLG 7.500,- per cursus. De kosten voor toetsing alleen bedroegen daarmee ruim NLG 300,- per deelnemer, ofwel NLG 40,- per deelnemer voor elke vaardigheid afzonderlijk. Voor experiment 2 en 3, waarbij toetsing onderdeel vormt van nascholing, en waarbij vier stations werden gebruikt, waren de kosten ongeveer NLG 400,- per deelnemer, ofwel ongeveer NLG 100,- per deelnemer voor elke vaardigheid afzonderlijk.

Tabel 3 Kostenoverzicht toetsing en training van technische vaardigheden

Kostenpost (in nederlandse guldens)	exp1	exp2	exp3
Materiaal voor toetsing/training	4.500	3.600	1.500
Kantoormiddelen en drukwerk	2.000 *	3.700	1.800
Reis- en transportkosten	1.900	300	450
Simulatiepatiënten	4.950	1.350	3.650
Observatoren (huisartsen)	7500 *	4.900	4.600
Trainers	n.v.t.	4.000	4.000
Personele ondersteuning	1.000 *	900	950
Organisatiekosten	7.500 *	7.500	7.500 *
Catering	2.200	2.200	1.800
Totale kosten	**31.550**	**28.450**	**26.450**
Aantal deelnemers	96	72	64
Kosten per deelnemer	328	395	413
Kosten per vaardigheid	41	98	103

* Cijfers op basis van schatting

Discussie

Uit de evaluatie gegevens kan opgemaakt worden dat (educatieve!) toetsing bij de deelnemende huisartsen weinig weerstanden oproept. Dat is bemoedigend, waarbij wel aangetekend moet worden dat de deelnemende huisartsen niet zonder meer beschouwd mogen worden als representatief voor de beroepsgroep. Immers zij waren van te voren op de hoogte dat zij aan diverse vormen van toetsing onderworpen zouden worden. Bij onderzoek onder huisartsen die

deelnamen aan nascholing in de regio Rotterdam bleek dat niet alle huisartsen enthousiast waren over dergelijke meer confronterende vormen van nascholing (Delnoy 1993). Een kwart van de respondenten gaf aan met gemengde gevoelens of liever niet deel te nemen aan een dergelijke vorm van nascholing. Ook in diverse artikelen over het kwaliteitsbeleid in de huisartsgeneeskundige pers klinken deze reserves door (Verdenius et al. 1990; Verdenius 1993). Uit een enquete-onderzoek onder 120 huisartsen, waarin onder meer gevraagd werd naar bekendheid met en toepassing van verschillende methoden van toetsing en kwaliteitsverbetering, bleek dat meer dan 80% toetsing van kennis en vaardigheden als een nuttige basis voor nascholing ervaarde. Echter ongeveer de helft van de respondenten vond dat zij onvoldoende bekend waren met de methode en slechts 11% had er ook daadwerkelijk ervaring mee. Angst voor toetsing (door anderen) bleek een belangrijke reden (Grol en Wensing 1995). Ook bij de deelnemers aan de bovengenoemde experimenten bestond overigens vooraf nog wel enige 'examenvrees', maar dit was in de enquête-resultaten (na afloop van de cursus ingevuld) niet terug te vinden.

De vaardigheidstoets geniet duidelijk meer waardering als toetsingsinstrument dan de kennistoets over vaardigheden. Blijkbaar vinden de deelnemers het nuttiger om getoetst te worden op hun beheersing van de vaardigheid dan alleen op de noodzakelijke kennis om vaardigheden toe te passen. De directere confrontatie met het eigen handelen die in vaardigheidstoetsing besloten ligt wordt in de educatieve setting niet bezwaarlijk gevonden. Voorts blijkt de waardering voor toetsing hoger indien toetsing plaats vindt als onderdeel van deskundigheidsbevordering in de vorm van een gerichte instructie van de te leren vaardigheden. Inbedding van toetsing in nascholing onderstreept de educatieve betekenis van toetsing, maakt het meer een onderdeel van het leren, terwijl een geisoleerde toets meer de examensituatie benadrukt.

Toetsing van technische vaardigheden kost, indien - afgezien van de kosten van materiaalontwikkeling - alle kosten worden doorberekend, NLG 40,- per vaardigheid. Indien training en toetsing volgens het beschreven model worden gecombineerd, dan zijn de kosten ongeveer NLG 100,- per vaardigheid, ofwel NLG 400,- per cursus. Deze bedragen zijn gebaseerd op de nederlandse omstandigheden en op basis van de geboden facilitaire ondersteuning vanuit de verschillende academische centra. Doorberekening van kosten op basis van externe dienstverlening zou naar schatting een kostenverhoging van 20 - 30 % betekenen. Daarmee zijn de kosten van vaardigheidstraining en toetsing in Nederland vergelijkbaar met opgaven in de buitenlandse literatuur (Reznick et al. 1993; Cusimano et al. 1994; Carpenter 1995). Ook in vergelijking met de kosten van reguliere vaardigheidstraining in Nederland zijn de meerkosten van toevoeging van toetsing bescheiden. Dit is verrassend, omdat vaardigheidstoetsing nogal eens als een (te) kostbare vorm van toetsing wordt bestempeld (Anderson en Kassebaum 1993). De goede acceptatie en relatief bescheiden kosten openen perspectieven voor toepassing van vardigheidstoetsing in reguliere nascholing van huisartsen. Voor implementatie worden echter ook eisen gesteld aan de beschikbaarheid van goede materialen en infrastructuur om dergelijke toetsingen op reguliere basis aan te kunnen bieden.

Samenvattend lijkt educatieve vaardigheidstoetsing ingebed in nascholing voor huisartsen een acceptabele toetsingsvorm, waarvan de potentiële meerwaarde ruimschoots opweegt tegen de geringe meerkosten.

Literatuur

Anderson MB, Kasssebaum DG (eds). Proceedings of the AAMC's consensus conference on the use of standardized patients in teaching and evaluation of clinical skills. Acad Med 1993;68:437-83.

Anoniem. Notitie 'Kwaliteit en Deskundigheidsbevordering'. Utrecht: LHV, 1990.

Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med 1995;70:828-33.

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE. Acad Med 1994;69:571-6.

Delnoy DMJ. Nascholing voor huisartsen in Rotterdam: een marktonderzoek. Utrecht: NIVEL, 1993.

Driessen SD. Studenten over vaardigheidsonderwijs: vaardigheden leren, hoe gaat dat? In: Dochy F, Van Luijk SJ (red). Handboek Vaardigheidsonderwijs. Lisse: Swets & Zeitlinger, 1987:303-13.

Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective structured clinical examination for licensing family physicians. Can Med Assoc J 1992;146:1735-40.

Grol R, Wensing M. Implementation of quality assurance and medical audit: general practitioners' perceived obstacles and requirements. Br J Gen Pract 1995;45:548-52.

Hays RB, Bridges-Webb C, Booth B. Quality assurance in general practice. Med Educ 1993;27:175-80.

Hays R. Methods of assessment in recertification. In: Newble D, Jolly B, Wakeford R (eds). The certification and recertification of doctors. Issues in the assessment of clinical competence. Cambridge: Cambridge University Press, 1994:187-200.

Jansen JJM, Tan LHC, Van der Vleuten CPM, Van Luijk SJ, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners. Med Educ 1995;29:247-53.

Jansen JJM, Scherpbier AJJA, Metz JCM, Grol RPTM, Van der Vleuten CPM, Rethans JJ. Construct validity of performance-based assessment in continuing medical education for general practitioners. Med Educ 1996;30:339-44.

Newble DI. Eight years' experience with a structured clinical examination. Med Educ 1988;22:200-4.

Pollemans MC, Tan LHC. Toetsing van Kwaliteit. Landelijke evaluatie van de interimbeoepsopleiding tot huisarts. Rapport SV-IOH-15. Utrecht: SV-IOH, 1990.

Reznick RK, Smee S, Rothman A, Chalmers A, Swanson D, Dufresne L, et al. An Objective Structured Clinical Examination for the Licentiate; Report from the Pilot Project of the Medical Council of Canada. Acad Med 1992;67:487-94.

Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, et al. Guidelines for estimating the real cost of an objective structured clinical examination. Acad Med 1993; 68: 513-7.

Verdenius W, Brands PJ, Oudkerk RH. Toetsing: killer of saviour? De Huisarts 1990;1(7):18-20.

Verdenius W. Huisarts en kwaliteitsbeleid (I). Wordt de individuele huisarts er wijzer van? De Huisarts 1993;4(3) :69-72.

Hoofdstuk 10

Beschouwing

Sinds de tijd dat het artsexamen nog een levenslang 'rijbewijs' voor het bedrijven van de huisartsgeneeskunde verschafte - en dat ligt nog maar net twee decennia achter ons - is er veel veranderd. De totstandkoming van een eigen opleiding (Runia en van Herk 1991) en een eigen onderzoekstraditie (Touw-Otten 1981) markeren de veranderingen, die geleid hebben tot de huidige positie van de huisarts in Nederland. De professionalisering van de huisartsgeneeskunde in Nederland en de veranderende verwachtingen vanuit de samenleving hebben impulsen gegeven tot de ontwikkeling van een kwaliteitssysteem, waarin de huisarts op allerlei manieren ondersteund wordt om zijn/haar zorg voor de patiënten te optimaliseren (Anoniem 1987; Grol 1991). Deskundigheidsbevordering op relevante aspecten van die zorg en regelmatige toetsing, om na te gaan in hoeverre doelstellingen worden verwezenlijkt, vormen de hoekstenen van dit kwaliteitsbeleid (Anoniem 1990; Grol 1993). Het is vooral aan de professie zelf om op dit gebied intern voor een goede kwaliteit te zorgen en die te bewaken (Anoniem 1995). De ontwikkeling van richtlijnen, nascholing en (onderlinge) toetsing vormen daartoe onmisbare schakels (Rutten en Thomas 1993; Thomas et al. 1996; Grol en Mesker 1986).
Het onderzoek waarover in dit proefschrift verslag is gedaan beoogt een bouwsteen aan te dragen voor de deskundigheidsbevordering en toetsing van technische vaardigheden van huisartsen. In dit hoofdstuk worden de verschillende aspecten van het onderzoek aan een nadere beschouwing onderworpen, om na te gaan wat de gebruikswaarde van een dergelijke bouwsteen zou kunnen zijn.

Medisch technische vaardigheden van huisartsen: domein en prioriteiten

De eerste vraagstellling van het onderzoek had betrekking op het domein van medisch-technische vaardigheden voor de huisarts. Uit de inventarisatie van het domein blijkt dat de Nederlandse huisarts een grote diversiteit aan diagnostische en therapeutische verrichtingen tot zijn arsenaal mag rekenen. Dit domein is geen statisch gegeven, maar verandert onder invloed technologische en maatschappelijke ontwikkelingen (Anoniem 1995). Regelmatige herziening van het takenpakket en nascholing van vaardigheden is mede daarom van belang voor de huisarts.
De prioriteiten die Nederlandse huisartsen stellen voor nascholing weerspiegelen de behoefte om bij te blijven met nieuwe ontwikkelingen, zoals op het gebied van de thuiszorg. Er ligt onder meer grote nadruk op vaardigheden die relevant zijn in het kader van de zorg voor ouderen en chronisch zieken. Naast kennismaking met nieuwe ontwikkelingen blijkt er echter ook een duidelijke behoefte om vaardigheden regelmatig op te frissen voor aandoeningen waarmee een huisarts veel te maken krijgt, zoals klachten van het bewegingsapparaat.

Methoden van toetsing van technische vaardigheden

De tweede vraagstelling betrof de geschiktheid van methoden voor toetsing van technische vaardigheden. Drie verschillende toetsingsmethoden zijn in dit onderzoek onderzocht op hun

merites: een vaardigheidstoets, en in vergelijking daarmee een kennistoets over vaardigheden, en een zelfbeoordelingslijst van vaardigheden.

Validiteit van vaardigheidstoetsing

In een eerste experiment werden de drie toetsingsmethoden toegepast bij een groep huisartsopleiders en huisartsen-in-opleiding. De verschillen in gemiddelde scores tussen beide groepen op de vaardigheidstoets en kennistoets waren gering, terwijl de meer ervaren huisartsen wel gemiddeld hoger scoorden op de zelfbeoordelingslijst. De geringe onderlinge verschillen in scores op de vaardigheidstoets voor verschillende opleidingsstadia en ervaring als huisarts roepen de vraag op in hoeverre de vaardigheidstoets valide is om verschillen in expertise van meer ervaren artsen te meten (Cox 1990; Norman et al. 1991). Waren er verschillen die het toetsings-instrument niet meet, of waren er geen verschillen? De scores op de kennistoets over vaardigheden lieten evenmin (grote) verschillen zien. Juist van kennistoetsen is bekend dat deze goed in staat zijn om verschillen in competentie te meten, ook bij meer ervaren artsen (Norman et al. 1994). De bevinding dat er geen verschillen waren tussen gemiddelde scores van huisartsen en huisartsen-in-opleiding voor zowel de kennistoets over vaardigheden als de vaardigheidstoets maakt het dus aannemelijker dat er geen sprake was van verschillen in vaardigheidsnivo.

In een tweede experiment werd de (construct)validiteit van vaardigheidstoetsing aan een nader onderzoek onderworpen. Toetsing maakte daarbij onderdeel uit van een nascholingscursus voor huisartsen. Bij toetsing bleek er een duidelijk verschil in score tussen de huisartsen die de training hadden gevolgd in vergelijking met de huisartsen die geen training hadden gevolgd. Blijkbaar waren de gebruikte instrumenten valide om verschillen in kennis- en vaardigheidsnivo vast te stellen.

De resultaten van deze twee experimenten maakten duidelijk dat zowel de kennistoets als vaardigheidstoets goed in staat waren om verschillen in competentie te meten. De in het eerste experiment gevonden kleine onderlinge verschillen in gemiddelde score tussen huisartsen-in-opleiding en ervaren huisartsen moeten dus inderdaad aan geringe verschillen in competentie worden toegeschreven, en niet aan een gebrek aan (construct) validiteit van de toetsings-instrumenten.

De betekenis van de geringe verschillen tussen huisartsen-in-opleiding en huisartsen is niet geheel duidelijk. De transversale onderzoeksopzet maakt interpretaties over veranderingen in de tijd van het gemiddelde vaardigheidsnivo moeilijk. Voor huisartsgeneeskundige kennis lijken wel veranderingen in kennisnivo aantoonbaar (Van Leeuwen et al. 1995).

Betrouwbaarheid van vaardigheidstoetsing

De betrouwbaarheid (dat wil zeggen generaliseerbaarheid) van de individuele vaardigheidstoetsscore bleek in het eerste experiment op basis van de acht vaardigheden waaruit de toets bestond, zoals verwacht, niet erg hoog (0.43). Deze resultaten zijn

vergelijkbaar met de resultaten in de internationale literatuur (Van der Vleuten en Swanson 1990), waarbij voor een toetsduur van 2 uur waarden gevonden worden tussen 0.31 en 0.70. Om echter een relatieve betrouwbaarheid (voor de rangordening van deelnemers) van 0.80 te bereiken, die in de literatuur vaak als ondergrens wordt gehanteerd (Dousma en Horsten 1980), zou getoetst moeten worden over 40 verschillende vaardigheden met een totale toetstijd van 10 uur. Voor een absolute betrouwbaarheid (ten aanzien van de hoogte van de score) van 0.80 van de score zou zelfs een toetstijd van 14,5 uur nodig zijn. Dergelijke toetstijden zijn in de praktijk niet haalbaar, en lijken ook niet wenselijk. Op basis van toetstijden van 2 uur, zoals in het eerste experiment, zijn echter wel betrouwbare beslissingen mogelijk, zoals bijvoorbeeld de beslissing of de deelnemer een voldoende beheersingsnivo van vaardigheden heeft, ondanks de beperkte betrouwbaarheid van de toetsscore (Van Luijk en Van der Vleuten 1988). Immers bij een dergelijke beslissing gaat het er niet zozeer om hoe precies het relatieve of absolute beheersingsnivo van de deelnemers is, maar of de score zich onder of boven de vastgestelde grenswaarde bevindt. Hoeveel erboven of eronder is niet van belang. Overigens is de betrouwbaarheid van de beslissing wel afhankelijk van de feitelijke score van de deelnemer. Hoe verder de score van de deelnemer verwijderd is van de vastgestelde grenswaarde, hoe groter de betrouwbaarheid van de beslissing (Van der Vleuten en Wijnen 1991).

Validiteit en betrouwbaarheid kennistoets en zelfbeoordelingslijst

De validiteit van de kennistoets en zelfbeoordelingslijst voor het vaststellen van vaardigheidsbeheersing werd vooral bepaald door de samenhang van kennistoets en zelfbeoordelingslijst met de vaardigheidstoets, waarop verderop wordt ingegaan. Zowel kennistoets als zelfbeoordelingslijst lieten in de diverse experimenten verschillen zien op groepsnivo die als ondersteunend voor de construct-validiteit beschouwd kunnen worden.

De kennistoets (met een norm-georienteerde betrouwbaarheid van 0.68 bij een toetstijd van 1 uur) en zelfbeoordelingslijst (norm-georienteerde betrouwbaarheid 0.92 bij invultijd van 10 minuten) bleken met veel kortere toetstijden aanvaardbare betrouwbaarheden op te leveren in vergelijking met de vaardigheidstoets, omdat met behulp van deze toetsen in korte tijd over een groot aantal verschillende vaardigheden toetsgegevens verzameld kunnen worden.

Samenhang tussen vaardigheidstoets en kennistoets

Bij beschouwing van de samenhang tussen de scores op de vaardigheidstoets en kennistoets blijken deze een sterke correlatie vertonen, indien gecorrigeerd voor onbetrouwbaarheid. Dit betekent dat een kennistoets over vaardigheden een redelijke voorspellende waarde heeft voor het vaardigheidsnivo van huisartsen. Dit verband werd ook bij ouderejaars medische studenten gevonden (Van der Vleuten et al. 1988; Newble en Swanson 1988). Aangezien allerlei toetsen van medische competentie onderling een vrije sterke samenhang in score laten zien (Van der Vleuten et al. 1994) is het echter de vraag in hoeverre een specifieke kennis over vaardigheden toets grote meerwaarde heeft boven een algemene kennistoets om het algemene

vaardigheidsnivo vast te stellen. Uit onderzoek (Van der Vleuten et al. 1988) onder medische studenten was de voorspellende waarde van een algemene kennistoets bijna even goed als die van een kennis over vaardighedentoets. Uit een eigen onderzoek onder huisartsen-in-opleiding bleek de specifieke kennis-over-vaardighedentoets wel een betere voorspelling te geven (Jansen et al. 1995).
De score op de kennistoets voor elk van de vaardigheden afzonderlijk bleek echter géén goede voorspeller te zijn voor de score op de vaardigheidstoets voor de betreffende vaardigheid. Ook in een andere studie werden dergelijke resultaten gevonden (Vu en Barrows 1990). Terwijl kennis en vaardigheidsbeheersing voor afzonderlijke vaardigheden een geringe samenhang vertonen, geldt dat er in het algemeen, dat wil zeggen uitgemiddeld over een groot aantal vaardigheden, wel een duidelijke samenhang tussen kennis en vaardigheidsbeheersing bestaat. Dit betekent niet dat een kennistoets en een vaardigheidstoets hetzelfde meten, maar bij toetsing over een groot aantal onderwerpen wel een vergelijkbare rangordening van deelnemers geven. Dit geldt overigens niet alleen voor kennis en vaardigheidsbeheersing, maar voor allerlei vormen van toetsing die medische competentie beogen te meten (Van der Vleuten et al. 1994).

Samenhang tussen vaardigheidstoetsing en zelfbeoordeling

Zelfbeoordeling heeft in vergelijking met kennistoetsing een lagere voorspellende waarde voor beheersing van technische vaardigheden. Opvallend was dat de meer ervaren huisartsen zichzelf als meer vaardig beoordeelden, terwijl de vaardigheidstoets en kennistoets hiervoor geen objectieve onderbouwing gaven. Dit zou kunnen betekenen dat deze toetsen relevante aspecten van vaardigheidsbeheersing niet meten (bijvoorbeeld routine en ervaring) die wel een rol spelen bij de zelfbeoordeling. Voor de diagnostische handelingen zou dit bijvoorbeeld kunnen betekenen dat ervaren huisartsen minder handelingen verrichten om tot een diagnose te komen op basis van hun expertise (Schmidt et al. 1990). Voor therapeutische verrichtingen is echter geen goede verklaring te bedenken voor de discrepantie. Een andere mogelijke verklaring is dat het algehele vertrouwen in eigen competentie toeneemt met ervaring, hetgeen ook doorwerkt in de zelfbeoordeling van vaardigheidsbeheersing als een soort 'halo-effect' (Streiner 1985). Op basis van de literatuur over zelfbeoordeling (Boud en Falchikov 1989; Gordon 1991; Gordon 1992) is de laatste verklaring waarschijnlijker. De lage voorspellende waarde van zelfbeoordeling voor beheersing van technische vaardigheden maakt aannemelijk dat de keuze voor nascholing van ervaren huisartsen op basis van zelfbeoordeling niet leidt tot keuze van onderwerpen die objectief gezien het meest voor de hand liggen, zoals ook al eerder werd gevonden (Sibley et al. 1982). Een recent onderzoek naar zelfbeoordeling van huisartsen ten aanzien van hun kennisnivo komt tot dezelfde bevinding (Tracey et al. 1997). Objectieve toetsing vormt daarom een noodzakelijk onderdeel voor ondersteuning van meer rationele keuze van nascholingsonderwerpen.

Het effect van toetsing op competentie en het handelen in de praktijk

De derde vraagstelling betrof de effectiviteit van nascholing en toetsing van technische vaardigheden op het handelen van de huisarts in de praktijk.

Het educatieve gebruik van toetsing in nascholing voor huisartsen heeft belangrijke voordelen. Het verschaft de deelnemers objectief inzicht in de eigen prestaties, en deze feedback lijkt de effectiviteit van de deskundigheidsbevordering te bevorderen (Grol 1992). In het tweede en derde experiment werd een duidelijk effect van de cursus op de competentie van de deelnemers vastgesteld, en dit effect was ook nog enkele maanden na de cursus duidelijk waarneembaar. In het derde experiment werd nagegaan in hoeverre een dergelijke 'meerwaarde' ook vast te stellen was in de vorm van effecten van de cursus op het handelen in de dagelijkse praktijk. Uit de literatuur is bekend dat competentie en feitelijk dagelijks handelen immers niet hetzelfde zijn (Rethans et al. 1991).Voor twee van de vier vaardigheden kon een effect in de praktijk gedemonstreerd worden, voor twee vaardigheden echter niet. De vaardigheden waarbij een effect in de praktijk optrad (schouderinjectie en cervix-uitstrijk), waren vaardigheden die de huisarts relatief frequent in de praktijk tegenkomt zodat het geleerde direct toegepast kon worden. Bovendien vereiste het toepassen van deze vaardigheden geen grote veranderingen in het handelen van de huisarts. Eén vaardigheid (fundoscopie bij diabetes mellitus) waarbij geen effect in de praktijk optrad was te moeilijk om in een korte cursus voldoende te leren beheersen, omdat de interpretatie van de bevindingen moeilijk is. Voor de andere vaardigheid (fluor diagnostiek) waren mogelijk te grote veranderingen in de routine van de praktijk nodig. Blijkbaar kan een korte cursus effectief zijn in het veranderen van het handelen in de praktijk. In de literatuur heerst er nogal wat scepsis over het nut van korte cursussen, en wordt veel nadruk gelegd op complexe interventies (Haynes et al. 1984; Davis et al. 1992; Wensing en Grol 1994).

De vorm van de nascholing - intensieve interactieve training met praktische oefening van vaardigheden in kleine groepen gecombineerd met toetsing en feedback - heeft mogelijk bijgedragen aan het effect op de praktijk. Ook anderen hebben positieve effecten op het praktisch handelen gevonden van nascholing die volgens duidelijke didactische principes is opgezet (Stein 1981). Er is dus een - zij het misschien bescheiden - plaats voor intensieve op het praktisch handelen gerichte nascholingscursussen, zoals gebruikt in dit onderzoek, als een effectief middel voor het tot stand brengen van veranderingen in het medische handelen in de dagelijkse praktijk. Er bestaat echter nog veel onduidelijkheid over hoe en wanneer nascholing effectief kan zijn (Kanouse et al. 1995). Mogelijk zijn meerdere strategiën effectief (Grol 1997).

In dit onderzoek zijn de verschillende toetsingsvormen steeds in combinatie gebruikt. Daardoor is het niet mogelijk uitspraken te doen over de afzonderlijke (en mogelijk verschillende) onderwijseffecten van vaardigheidstoetsing, kennistoetsing en zelfbeoordeling. Uit de literatuur is bekend dat toetsing een sterke invloed heeft op het leren, waarbij de toetsvorm ook de vorm van het leren beïnvloed. Vaardigheidstoetsing stimuleert leren gericht op vaardigheids-

beheersing, terwijl kennistoetsing het verwerven van kennis stimuleert (Newble en Jaeger 1983; Frederiksen 1984; Stillman en Swanson 1987). Bij de keuze voor een bepaalde toetsvorm zullen, naast de psychometrische eigenschappen, deze educatieve effecten van verschillende toetsvormen in de beschouwing betrokken dienen te worden om een optimaal rendement van toetsing te verkrijgen.

De invloed van feedback op de zelfbeoordeling van huisartsen

Behalve de effecten van nascholing en toetsing op het handelen van de huisarts werd in het derde experiment ook de invloed van feedback op de zelfbeoordeling van huisartsen onderzocht. Hieraan lag de veronderstelling ten grondslag dat zelfbeoordeling een vaardigheid is die geleerd moet worden (Gordon 1992). Zelfbeoordeling bleek geen goede voorspelling te geven van het vaardigheidsnivo en er was ook geen duidelijk leereffect merkbaar. Deze bevindingen onderstrepen nog eens het belang om keuzen voor nascholing van huisartsen zoveel mogelijk te baseren op objectieve (toets) gegevens.

Waardering en kosten van toetsing

De vierde onderzoeksvraag betrof de haalbaarheid van educatieve vaardigheidstoetsing van huisartsen. Om nascholingsvormen met toetsing te kunnen implementeren dient rekening gehouden te worden met de bereidheid van huisartsen zich aan toetsing te onderwerpen en met de kosten.

Uit de enquetes die werden afgenomen blijkt dat huisartsen niet onwelwillend tegenover educatieve toetsing staan, vooral indien toetsing onderdeel uitmaakt van nascholing. Voor toetsing als geïsoleerde activiteit lijkt het enthousiasme geringer. De vaardigheidstoets heeft de voorkeur boven de schriftelijke toets. De toetsing en feedback op vaardigheidsbeheersing wordt nuttiger gevonden in vergelijking met de toetsing op kennis. De observatie en directe confrontatie met onvolkomenheden in de beheersing roepen - in de educatieve setting van de nascholing - nauwelijks weerstanden op. De vraag is echter of deze welwillende houding kenmerkend is voor de beroepsgroep als geheel. De deelnemers aan de experimenten vormen immers een selectie van huisartsen die bereid was zich toetsbaar op te stellen. Toetsing roept bij veel huisartsen onaangename associaties met vroegere examen-situaties op (Delnoy 1993), en dus ook weerstand (Grol en Wensing 1995), hetgeen ook doorklinkt in artikelen over kwaliteitsbeleid in de huisartsgeneeskundige pers (Verdenius et al. 1990; Verdenius 1993). Door op uitgebreidere schaal huisartsen kennis te laten maken met toetsing in het kader van deskundigheidsbevordering kunnen deze weerstanden op den duur mogelijk plaatsmaken voor meer enthousiasme.

De kosten van vaardigheidstraining inclusief toetsing blijken in verhouding tot reguliere nascholing heel acceptabel, en liggen in de orde van NLG 100,- per vaardigheid, waarvan de toetsingscomponent ongeveer NLG 40,- bedraagt. In vergelijking met de literatuur (Reznick et al. 1993; Cusimano et al. 1994; Carpenter 1995), waar kosten van toetsing NLG 65-100,-

per vaardigheid worden genoemd, zijn de kosten in Nederland relatief laag. De kosten hoeven dan ook geen belemmering te vormen om periodieke toetsing in te voeren als onderdeel van geaccrediteerde nascholing.

Methodologische kanttekeningen

Domeinbeschrijving en prioriteitstelling

De samenstelling van een lijst van medisch-technische vaardigheden op basis van een aantal bronnen bleek geen eenvoudige opgave. De ongelijksoortigheid van vaardigheden op de bestaande lijsten vormde een eerste probleem bij het samenstellen van de lijst. Sommige vaardigheden waren sterk opgesplitst in deelvaardigheden, zoals bijvoorbeeld het 'onderzoek van de schouder', dat in één lijst was opgesplitst in een tiental deelaspecten, zoals 'actief en passief onderzoek nek', 'actief en passief onderzoek schouder', 'bewegingsonderzoek schouder', 'inspectie schouder', 'weerstandtests schouderbewegingen'. Anderzijds werden complexe vaardigheden onder een algemene noemer verenigd, zoals 'begeleiding baring'. Ook werden vaardigheden onder een gemeenschappelijke algemene noemer gebracht, zoals 'inspectie gewrichten', die vanuit het perspectief van het probleemgericht of klachtgericht werken van huisartsen geen goede basis bieden voor toetsing en/of nascholing. Enerzijds door samenvoeging van een aantal deelvaardigheden, anderzijds door het onderbrengen van algemene categorieën bij verschillende vaardigheden, werd getracht de onevenwichtigheid van de lijst enigszins verminderd. Deze procedure was echter niet vrij van onderzoekers bias, en de vraag is of een ander niet afwijkende keuzen zou hebben gemaakt, resulterend in een andere lijst. De classificatie van vaardigheden vroeg eveneens om arbitraire beslissingen. De gebruikelijke classificaties, zoals de ICPC, zijn namelijk probleem-georienteerd en niet vaardigheid-georienteerd. Dit betekent dat dezelfde vaardigheden bij diverse problemen van toepassing kunnen zijn. Overigens viel het aantal vaardigheden dat op nivo van hoofdstukken van de ICPC niet eenduidig ingedeeld kon worden uiteindelijk mee.
Een tweede probleem vormde de eerste selectie van vaardigheden die prioriteit zouden moeten krijgen voor nascholing. Er werden een aantal criteria geformuleerd, op basis waarvan de onderzoeker een selectie maakte. De criteria bleken echter moeilijk te operationaliseren tot een goed hanteerbaar selectie instrument, zodat de onderzoeker als zodanig fungeerde.
De uiteindelijke lijsten die in het onderzoek zijn gebruikt zijn middels een pragmatische aanpak tot stand gekomen. Nadere validering zal echter moeten uitmaken of de lijsten ook in ander verband bruikbaar zijn.

De normstelling bij vaardigheidstoetsing

De resultaten op de verschillende toetsen suggereren dat de beheersing van technische vaardigheden van huisartsen niet in alle opzichten optimaal is. Dit is ook uit ander onderzoek is gebleken (Campbell et al. 1991; Fisher en Pfleiderer 1992; Berden 1993; Reenders et al.

1992). Toch ontbreekt het aan de noodzakelijke informatie om duidelijke conclusies over de kwaliteit van het handelen te trekken. Voldoende beheersing veronderstelt een norm. De scoringslijsten die gebruikt werden tijdens de vaardigheidstoets bevatten aspecten die van belang worden geacht voor een goede uitvoering van de vaardigheid. Het is echter de vraag of al die aspecten even relevant of essentieel zijn voor een goede vaardigheidsbeheersing. De scoringslijsten weerspiegelen niet zozeer een noodzakelijk minimum alswel een optimum, en dus meer een ideaal dan een norm met hoge realiteitswaarde. Het is niet verwonderlijk dat de gemiddelde score van werkelijke huisartsen daar onder valt (Donabedian 1982). Dit nivo van vaardigheidsbeheersing kan voldoende zijn om in de dagelijkse praktijk goede zorg te leveren. In het onderzoek is niet onderzocht wat een goede norm zou zijn, noch hoe die tot stand zou moeten komen. Er zijn allerlei varianten van benaderingen om tot een norm te komen (Livingston en Zieky 1982; Van Luijk en Wijnen 1996). Vanuit een educatieve benadering van toetsing - en die benadering vormde het uitgangspunt bij dit onderzoek - zijn dergelijke normen van relatief belang. Het leren staat dan voorop, en daarbij is het hanteren van een hoog streefnivo onderdeel van het leren. Bij selectieve toepassing van toetsing geldt daarentegen doorgaans als norm het minimaal aanvaardbare beheersingsnivo. Dit verandert wel de context van toetsing. Het toetsen dient dan niet meer (uitsluitend) het individuele belang van de huisarts-deelnemer, maar vooral een collectief belang.

Betrouwbaarheid van observatoren

De betrouwbaarheid van beoordeling door observatoren is in het algemeen vrij hoog bij toetsing van technische vaardigheden (Van der Vleuten en Swanson 1990). Toch kunnen observatoren onderling nogal van mening verschillen (Wakefield 1985) en ook systematische bias vertonen (Van der Vleuten et al. 1989). Als onderdeel van het tweede experiment kon de beoordeling door observatoren nader beschouwd worden voor basale reanimatie technieken. Omdat voor reanimatie zowel een door huisartsen ontwikkelde scoringslijst beschikbaar was als een recentelijk ontwikkeld mechanisch scoringssysteem (Berden 1993), was er de mogelijkheid om beide scoringssystemen te vergelijken, en daarmee ook een nadere indruk te verkrijgen omtrent de nauwkeurigheid van de observator. Daarbij bleek dat met name het observeren van de reanimatie handelingen minder betrouwbaar was, terwijl voor de diagnostische handelingen wel een redelijk betrouwbare beoordeling werd verkregen. Deze resultaten vormen een ondersteuning voor het gebruik van een combinatie van een scoringslijst op basis van observatie (voor de diagnostiek) en mechanische registratie (voor de reanimatie handelingen).

Meer in het algemeen relativeren deze resultaten de nauwkeurigheid van de observator voor verschillende onderdelen van een scoringslijst. Bepaalde handelingen zijn blijkbaar moeilijker te observeren. Observatortraining helpt weliswaar in het algemeen om de eenduidigheid in het beoordelen te verbeteren, maar juist bij professionals is het effect van observator-training relatief gering (Van der Vleuten et al. 1989; Newble et al. 1980). Voor educatieve toetsing is

dat niet zo bezwaarlijk. Voor eventuele selectieve toepassing dient wel met dit gegeven rekening te worden gehouden. Er zijn overigens wel mogelijkheden om dat observator-effect te neutraliseren. Bij een algemene toets (met een grote hoeveelheid stations) middelen de effecten van de verschillende observatoren uit, indien voor elk station andere observatoren worden gebruikt. Bij toetsing van één onderwerp zou met meerdere observatoren gewerkt kunnen worden (Van der Vleuten en Wijnen 1991). Voorts kunnen ook minder betrouwbare beoordelingen een basis vormen voor betrouwbare beslissingen. In selectieve toetssituaties gaat het immers niet zozeer om de betrouwbaarheid van de score alswel om de betrouwbaarheid van de beslissing of de deelnemer al of niet voldoende kennis danwel vaardigheidsbeheersing heeft. Voor dergelijke zak/slaag beslissingen kan, afhankelijk van de grenswaarde ten opzichte van de scores van de deelnemers, de betrouwbaarheid van de zak/slaag beslissing hoog zijn, ondanks een beperkte betrouwbaarheid van de toetsscore (Van der Vleuten en Van Luijk 1988).

Generaliseerbaarheid van onderzoeksresultaten

In dit proefschrift worden de mogelijkheden van toetsing van vaardigheden in het kader van deskundigheidsbevordering van huisartsen beschreven. Op basis van de experimenten waarbij een relatief beperkt aantal onderwerpen werd getoetst, bij een selecte groep huisartsen die bereid waren zich aan toetsing te onderwerpen, kunnen conclusies en aanbevelingen niet zonder meer worden gegeneraliseerd. De resultaten staan echter niet op zichzelf, en vinden bevestiging in wat bij andere groepen aan resultaten is gerapporteerd.

In de diverse experimenten is getracht om 'bias' zoveel mogelijk te voorkomen. Hier en daar is echter uitval opgetreden. Er zijn geen aanwijzingen gevonden dat er sprake was van selectieve uitval, maar methodologisch vormt het een zwak punt. Het is echter wel de realiteit van onderzoek onder praktizerende huisartsen die vele verschillende verplichtingen en soms conflicterende prioriteiten hebben. Medewerking aan het onderzoek is er slechts één van. Anderzijds vormt het verrichten van onderzoek in situaties die zo nauw mogelijk aansluiten bij de werkelijkheid van de doelgroep een betere waarborg voor toepasbaarheid dan sterk gecontroleerde experimenten die de doelgroep als 'vreemd' ervaart.

Aanbevelingen voor verder onderzoek

Hiervoor is aangegeven dat het onderzoek een aantal beperkingen heeft gekend, waardoor de bevindingen mogelijk niet in alle opzichten een algemenere geldigheid hebben. Het zou daarom wenselijk zijn om vaardigheidstoetsing te herhalen met andere onderwerpen en andere (liefst grotere groepen) deelnemers om na te gaan of dit tot vergelijkbare uitkomsten leidt. Daarbij zou verder onderzoek naar de veranderingen in vaardigheidsbeheersing gedurende de huisartsopleiding en onder practizerende huisartsen inzicht kunnen geven in de behoefte aan vaardigheidstraining. De effectiviteit van korte vaardigheidscursussen om veranderingen te bewerkstelligen in de praktijk is bemoedigend voor betrokkenen bij nascholing. Nascholing,

georganiseerd volgens doordachte didactische principes, kan wel degelijk effect sorteren. Niettemin is ook duidelijk dat er nog veel onbekend is over wanneer en hoe en bij wie dit effect optreedt (Kanouse et al. 1995, Grol 1997). Welke interventies de grootste effectiviteit hebben, lijkt sterk af te hangen van het onderwerp. Er is behoefte aan enerzijds een verdere theoretische fundering van het leren en veranderen van artsen in hun professionele leven, en anderzijds empirisch onderzoek naar de effectiviteit van verschillende interventies (Fox en Davis 1994; Kanouse et al. 1995; Grol 1996), waarbij intuitie en traditie worden vervangen door een rationele 'evidence based' benadering, zoals ook voor het medische basiscurriculum wordt voorgestaan (Van der Vleuten 1996). Het toetsingsinstrumentarium, zoals ontwikkeld voor de studies in dit proefschrift, zou daarbij voor het vaardighedendomein van huisartsen als bruikbaar hulpmiddel kunnen dienen.

Toetsing is in dit onderzoek steeds educatief gebruikt, en vormt in dat opzicht ook een krachtig onderwijsinstrument. De vraag naar de selectieve betekenis van toetsing is buiten beschouwing gebleven. Elders is daar wel al ervaring mee opgedaan, zowel in het basiscurriculum (Van Luijk en Van der Vleuten 1992; Vu et al. 1992) als in de huisartsopleiding (Grand'Maison et al. 1992), en op experimentele basis bij practizerende artsen (Norman et al. 1993; Caulford et al. 1994). Het is de vraag in hoeverre toepassing van selectieve toetsing in de Nederlandse huisartsgeneeskunde mogelijk en gewenst is. Hiermee verband houden vragen over wat een voldoende vaardigheidsbeheersing is voor huisartsen, ofwel een discussie over de normstelling (Van Luijk en Wijnen 1996; Cusimano 1996; Norcini 1992). De vragen over de implementatie zijn eveneens van groot belang, mede gezien het beladen karakter van toetsing (Delnoy 1993; Verdenius et al. 1990; Verdenius 1993). Voorts is ook niet vanzelfsprekend dat met toetsing de beoogde kwaliteitsverbetering wordt bereikt (Berwick 1992). Het belang van het ontwikkelen van normering voor vaardigheidsbeheersing is echter ook onderwijskundig gezien van groot belang, omdat het de scores op de vaardighedenstations, behalve een relatieve betekenis ('hoe is de score van de deelnemer in vergelijking met de overige deelnemers) ook een meer absolute betekenis geeft ('hoe is de score van de deelnemer in vergelijking met de norm'). Een nader onderzoek naar de normstelling zou bijvoorbeeld in eerste instantie kunnen geschieden in het kader van de huisartsopleiding, ter vaststelling of huisartsen aan het eind van hun opleiding tenminste over een vastgesteld minimumnivo van vaardigheidsbeheersing beschikken, zoals elders reeds gebeurt (Grand'Maison et al. 1992).

Aanbevelingen voor de praktijk

In het voorafgaande zijn al verschillende aanbevelingen gedaan voor praktische toepassing van toetsing van technische vaardigheden. Wat betreft de methoden die zijn onderzocht is vaardigheidstoetsing door middel van directe observatie de meest geschikte methode voor individuele educatieve toetsing van afzonderlijke vaardigheden. Voor individuele toetsing van het algemene vaardigheidsnivo vormt een kennistoets over vaardigheden een bruikbaar alternatief voor een vaardigheidstoets, omdat de kennistoets een redelijke voorspellende waarde

heeft en eenvoudiger is toe te passen dan een vaardigheidstoets (figuur 1). Bezien vanuit het leereffect van de toets heeft de vaardigheidstoets echter de voorkeur. Zelfbeoordeling is niet geschikt als methode voor individuele educatieve toetsing. De toetsing dient bij voorkeur onderdeel te vormen van nascholing. Voor toetsing op groepsnivo, bijvoorbeeld voor het vaststellen van nascholingsbehoefte of effecten van nascholingsprogramma's, lijken zowel een vaardigheidstoets als een kennistoets bruikbare methoden. In zeker mate is ook zelfbeoordeling geschikt om op groepsnivo veranderingen in vaardigheidsbeheersing vast te stellen. Ook hier zal de afweging gemaakt moet worden tussen enerzijds het leereffect van de toetsvorm en de praktische haalbaarheid.

Figuur 1 Bruikbaarheid van diverse methoden voor toetsing van technische vaardigheden

	individueel		groepsnivo	
	per vaardigheid	algemeen	per vaardigheid	algemeen
Vaardigheidstoets	+ + +	+ +	+ +	+ +
Kennistoets over vaardigheden	-	+ +	+ +	+ +
Zelfbeoordelingslijst	-	-	+	+

Resultaten van onderzoek leiden niet altijd vanzelfsprekend tot praktische toepassing, en dat geldt zeker ook voor het onderwerp van het onderzoek waarover is gerapporteerd. 'Onbekend maakt onbemind' lijkt ook voor vaardigheidstoetsing op te gaan (Grol en Wensing 1995). Eerder is al gewezen op het belang van acceptatie van toetsing, en het laten deelnemen van huisartsen aan vaardigheidstoetsing lijkt een goede manier om de acceptatie te bevorderen. Dit zal een belangrijk aandachtspunt moeten zijn van elke implementatie strategie. De welwillende houding van de Nederlandse huisartsen ten aanzien van instrumenten voor kwaliteitsverbetering en de goede acceptatie van educatieve toetsing door de deelnemers aan de experimenten in dit proefschrift vormen echter een geschikte basis om educatieve en screenende toetsing op ruimere schaal te introduceren.

Wat betreft de noodzakelijke randvoorwaarden zijn verbeteringen mogelijk. Weliswaar kan vaardigheidstoetsing in principe op velerlei locaties plaatsvinden, maar het is zeer wenselijk om te beschikken over goede voorzieningen, wat betreft mensen, ruimte en materialen, wil men met vaardigheidstoetsing verder komen dan het experimentele stadium. De meeste universitaire centra beschikken inmiddels over geschikte voorzieningen en ervaring met organisatie van training en toetsing van vaardigheden. Daarnaast zouden mogelijk ook nog enkele andere locaties een dergelijke centrumfunctie kunnen krijgen, zodat een goede

geografische spreiding over het land wordt verkregen. Door samenwerking tussen artsopleiding, huisartsopleiding en nascholingsorganisatie zou optimaal van deze voorzieningen gebruik gemaakt moeten worden, zodat voor elke huisarts dergelijke faciliteiten binnen redelijke afstand beschikbaar zijn.

Deze centra dienen materiaal ter beschikking te hebben voor training en toetsing. Gebleken is dat het ontwikkelen van dit materiaal een arbeidsintensief proces is. Het ligt dan ook voor de hand om ook hierin de krachten te bundelen en landelijk samen te werken. Net zoals de NHG-standaarden beogen om een groot deel van het huisartsgeneeskundige handelen te dekken, zouden de vaardighedentraining en toetsing als een uitwerking daarvan beschouwd kunnen worden voor het vaardighedendomein van de huisarts.

De kosten van vaardigheidstoetsing en vaardigheidstraining komen overeen met die van reguliere nascholing. De ontwikkelingskosten zijn daarin echter niet meegenomen. Financiering van dergelijke kosten zou overeenkomstig financiering van standaarden en deskundigheidsbevorderingspakketten kunnen geschieden.

De acceptatie verdient ook nadrukkelijk aandacht. De jongere generatie huisartsen zullen in hun opleiding vertrouwd worden gemaakt met periodieke toetsing van hun handelen, hetgeen op den duur ook effect zal hebben op de beroepsgroep. Daarnaast zijn echter ook activiteiten gericht op acceptatie in de beroepsgroep gewenst.

De huidige accreditering berust vooral op deelname aan nascholing die educatief aan bepaalde eisen voldoet. Met het verschijnen van bruikbare toetsingsinstrumenten op diverse deelaspecten van de huisartsgeneeskundige competentie, lijkt het redelijk om in de accreditering ook een zwaarder accent te leggen op toetsing, vanuit de gedachte dat deskundigheid zo mogelijk ook moet blijken. Deelname aan educatieve toetsing zou daartoe aantrekkelijk gemaakt moeten worden. De uitdaging voor de komende jaren is om de verschillende vormen van toetsing - waaronder vaardigheidstoetsing - in de nascholingspraktijk te integreren.

Literatuur

Anoniem. De positie van de huisarts in de toekomst. Utrecht: LHV, 1987.

Anoniem. Kwaliteits- en deskundigheidsbevordering. Utrecht: LHV, 1990.

Anoniem. De wereld verandert en de huisarts verandert mee.Utrech: LHV, 1995.

Berden HJJM. Basic Cardiopulmonary Resuscitation. Assessment of skills in training situations. (Proefschrift). Utrecht: Universiteit van Utrecht, 1993.

Berwick DM. Heal thyself or heal thy system: can doctors help to improve medical care? Quality in Health Care 1992;1:S2-8.

Boud D, Falchikov N. Quantitative studies of student-self-assessment in higher education: a critical analysis of findings. Higher Educ 1989;18:529-49.

Campbell HS, Fletcher SW, Lin S, Pilgrim CA, Morgan TM. Improving physicians' and nurses clinical breast

examination: a randomized controlled trial. Am J Prev Med 1991;7:1-8.

Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med 1995;70:828-33.

Caulford PG, Lamb SB, Kaigas TB, Hanna E, Norman GR, Davis DA. Physician incompetence: specific problems and predictors. Acad Med 1994;69:S16-8.

Cox K. No Oscar for OSCA. Med Educ 1990;24:540-5.

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. A comparative analysis of the costs of administration of an OSCE. Acad Med 1994;69:571-6.

Davis DA, Thomson MA, Oxman AD, Haynes B. Evidence for the effectiveness of CME. A review of 50 randomized controlled trials. JAMA 1992;268:1111-7.

Delnoy DMJ. Nascholing voor huisartsen in Rotterdam: een marktonderzoek. Utrecht: NIVEL, 1993.

Dousma T, Horsten A. Tentamineren. Groningen: Wolters-Noordhoff, 1980.

Fisher EW, Pfleiderer AG. Assessment of otoscopic skills of general practitioners and medical students: is there room for improvement? Br J Gen Pract 1992;42:65-7.

Frederiksen N. The real test bias: influences of testing on teaching and learing. Am Psychol 1984;39:193-202.

Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 1991;66:762-9.

Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med 1992;67:672-9.

Grand'Maison P, Lescop J, Rainsberry P, Brailovsky CA. Large-scale use of an objective strucured clinical examination for licensing family physicians. Can Med Assoc J 1992;46:1735-40.

Grol RPTM. Naar een 'kwaliteitssysteem' in de huisartsgeneeskunde.(Inaugurale rede). Utrecht: NHG, 1991.

Grol RPTM, Mesker PJR. Huisarts en onderlinge toetsing. Utrecht: Bunge, 1986.

Grol R. Implementing guidelines in general practice. Quality in Health Care 1992;1:184-91.

Grol R. Kwaliteitssystemen in de huisartsgeneeskunde: wat betekent dit voor de huisarts? Huisarts Wet 1993;36:106-12.

Grol R, Wensing M. Implementation of quality assurance and medical audit: general practitioners' perceived obstacles and requirements. Br J Gen Pract 1995;45:548-52.

Grol R. Beliefs and evidence in changing clinical practice. Br Med J 1997;315:418-21.

Haynes RB, Davis DA, McKibbon A, Tugwell P. A critical appraisal of the efficacy of continuing medical education. JAMA 1984;251:61-4.

Jansen JJM, Eekhof JAH, Dusman H. The predictive validity of a written test to assess competence of technical clinical skills in general practice. In:Rothman AI, Cohen R (eds). Proceedings of the sixth Ottawa conference on medical education. Toronto: University of Toronto, 1995:393-5.

Kanouse D, Kallich J, Kahan J. Dissemenination of effectiveness and outcomes research. Health Policy 1995;34:167-92.

Livingston SA, Zieky MJ. Passing Scores. Princetown: Educational Testing Service, 1982.

Newble DI, Hoare J, Sheldrake PF. The selection and training for clinical examinations. Med Educ 1980;14:345-9.

Newble D, Jaeger K. The effect of assessments and examinations on the learning of medical students. Med Educ 1983;17:165-71.

Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ

1988;23:325-34.

Norcini J. Approaches to standard-setting for performance-based examinations. In: HardenRM, Hart IR, Mulholland H (eds). Approaches to the assessment of clinical competence. Dundee, Centre for Medical Education, 1992:32-7.

Norman GR, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Med Educ 1991;25:119-26.

Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. JAMA 1993;270:1046-51.

Norman GR, Trott AD, Brooks LR, Smith EKM. Cognitive differences in clinical reasoning related to postgraduate training. Teaching and Learning in Medicine 1994;6:114-20.

Patrick J. Training: Research and Practice. London: Academic Press, 1992.

Reenders K, De Nobel E, Van den Hoogen HJM, van Weel C. Screening for diabetic retinopathy by general practitioners. Scand J Primary Health Care 1992;10:306-9.

Rethans JJ, Sturmans F, Drop R, Van der Vleuten C, Hobus P.Does competence of general practitioners predict their performance. Br Med J 1991;303:1377-85.

Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, et al. Guidelines for estimating the real cost of an objective structured clinical examination. Acad Med 1993;68:513-7.

Runia E, Van Herk R. De kunst van het haalbare. De verwezenlijking van de beroepsopleiding tot huisarts 1956-1973. Huisarts Wet 1991;34:117-23.

Rutten GEHM, Thomas S (red). NHG-standaarden voor de huisarts. Utrecht: Bunge, 1993.

Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. Acad Med 1990;65:611-21.

Sibley JC, Sackett DL, Neufeld V, Gerrard B, Rudnick KV, Fraser W. A randomized trial of continuing medical education. N Eng J Med 1982;306:511-5.

Stillman P, Swanson D. Ensuring the clinical competence of medical school graduates through standardized patients. Arch Intern Med 1987;147:1049-52.

Streiner DL. Global rating scales. In: Neufeld VR, Norman GR (eds). Assessing Clinical Competence. New York: Springer, 1985:119-41.

Tan LHC. Tekorten in de opleiding van huisartsen. ziektebeelden en medisch-technische vaardigheden. (Proefschrift). Amsterdam: Universiteit van Amsterdam, 1989.

Thomas S, Geijer RMM, Van der Laan JR, Wiersma T. NHG-standaarden voor de huisarts II. Utrecht: Bunge, 1996.

Touw-Otten FWMM. Wetenschapsbeoefening en huisartsgeneeskunde. Een analyse van dissertaties en enkele wegen tot structurering van huisartsgeneeskunde als discipline. (Proefschrift). Utrecht: Rijksuniversiteit Utrecht, 1981.

Tracey JM, Arroll B, Richmond DE, Barham PM. The validity of general practitioners' self-assessment of knowledge: cross sectional study. Br Med J 1997;315:1426-8.

Van der Vleuten CPM, Van Luijk SJ, Beckers HJM. A written test as an alternative to performance testing. Med Educ 1988;22:97-107.

Van der Vleuten CPM, Van Luijk SJ. Betrouwbaarheid van observatietoetsen voor praktische vaardigheden in het medisch onderwijs. Tijdschr Onderwijsres 1988;13:213-26.

Van der Vleuten CPM. Beyond Intuition (Inaugurale rede). Maastricht: Universitair Pers Maastricht, 1996.

Van der Vleuten CPM, Van Luijk SJ, Van Ballegooijen AMJ, Swanson DB. Training and experience of examiners. Med Educ 1989; 23: 290-6.

Van der Vleuten CPM, Wijnen WHFW. Niets praktischer dan een goede theorie: generaliseerbaarheidstheorie als instrument voor betrouwbaarheidsstudies. Bulletin Medisch Onderwijs 1991; 10: 2-14.

Van der Vleuten C, Newble D. Methods of assessment in certification. In: Newble D, Jolly B, Wakeford R (eds). The certification and recertification of doctors. Issues in the assessment of clinical competence. Cambridge:Cambridge University press, 1994:105-25.

Van der Vleuten CPM, Swanson DB. Assessment of skills with standardized patients: state of the art. Teach Learn Med 1990;2:58-76.

Van Leeuwen YD, Mol SSL, Pollemans MC, Drop MJ, Grol R, Van der Vleuten CPM. Change in knowledge of general practitioners during their professional careers. Fam Pract 1995;12:313-7.

Van Luijk SJ, Van der Vleuten CPM. A comparison of standard setting methods applied to a performance-based test. In: HardenRM, Hart IR, Mulholland H (eds). Approaches to the assessment of clinical competence. Dundee, Centre for Medical Education, 1992:326-30.

Van Luijk SJ, Wijnen WHFW. Cesuurbepaling. In: Metz JCM, Scherpbier AJJA, Van der Vleuten CPM (red). Medisch onderwijs in de praktijk. Assen: Van Gorcum, 1995:238-46.

Verdenius W, Brands PJ, Oudkerk RH. Toetsing: killer of saviour? De Huisarts 1990;1(7):18-20.

Verdenius W. Huisarts en kwaliteitsbeleid (I). Wordt de individuele huisarts er wijzer van? De Huisarts 1993;4(3):69-72.

Vu NV, Barrows HS. Validity and accuracy of performance and written evaluations in assessing history and physical examination skills. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds). Teaching and Assessing Clinical Competence. Groningen: Boekwerk, 1990:283-7.

Vu NV, Barrows HS, Marcy ML, Verhulst SJ, Colliver JC, Travis TA. Six years of comprehensive, clinical performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. Acad Med 1992;67:42-50.

Wakefield J. Direct observation. In: Neufeld VR, Norman GR (eds). Assessing clinical competence. New York, Springer, 1985:51-70.

Wensing M, Grol R. Single and combined strategies for implementing changes in primary care: a literature review. Quality in Health Care 1994;6:115-32.

Wensing M. Patients evaluate general practice. (Proefschrift). Nijmegen, 1997.

Samenvatting

Technische vaardigheden vormen belangrijk gereedschap bij het dagelijkse handelen van de Nederlandse huisarts. Onder medisch-technische vaardigheden worden in dit proefschrift vaardigheden verstaan die de huisarts gebruikt bij het uitvoeren van patiëntgebonden diagnostische of therapeutische handelingen. In concreto gaat het daarbij om lichamelijk onderzoek, om aanvullend onderzoek in de vorm van functietests of laboratoriumonderzoek, en om therapeutische ingrepen. In de studies die in dit proefschrift beschreven worden, is onderzocht met welke methoden en instrumenten de deskundigheid van huisartsen op het gebied van technische vaardigheden getoetst kan worden in het kader van de deskundigheidsbevordering van huisartsen.

In hoofdstuk 1 worden de achtergronden van het onderzoeksproject geschetst. De professionalisering van de huisartsgeneeskunde en maatschappelijke behoefte aan meer systematische aandacht voor de kwaliteit van de huisartsgeneeskunde, hebben de basis gelegd voor de ontwikkeling van een kwaliteitssysteem voor de huisartsgeneeskunde. In dat verband worden richtlijnen voor effectieve zorg, toetsingsmethoden om vast te stellen in hoeverre de feitelijke zorg daarmee overeenstemt, en methoden voor kwaliteitsverbetering ontwikkeld. Technische vaardigheden zijn een van de aandachtsgebieden.
Er blijkt weinig informatie te zijn over de feitelijke vaardigheidsbeheersing van praktizerende huisartsen voor wat betreft de vaardigheden uit het Basistakenpakket. Bovendien hebben zich sinds het Basistakenpakket is vastgesteld ontwikkelingen voorgedaan, waardoor wellicht herziening van het pakket gewenst is.
Een tweede terrein waarop onduidelijkheid bestond gold de geschiktheid van diverse methoden van toetsing voor het meten van vaardigheidsbeheersing van praktizerende huisartsen.
Vanuit de noodzaak om toetsing in nascholing te integreren, was een derde punt van aandacht of toetsing ook de effectiviteit van nascholing ten goede zou komen. In het bijzonder waren we daarbij geïnteresseerd in de mogelijkheden en effecten van feedback naar de deelnemers.
Tot slot moest het onderzoek meer duidelijkheid opleveren over de acceptatie en kosten van vaardigheidstoetsing en de organisatorische randvoorwaarden die daarbij gewenst zijn, teneinde een beter beeld te krijgen van de toepasbaarheid.

De volgende vraagstellingen werden geformuleerd:
1. *Welke medisch-technische vaardigheden vallen binnen het domein van de Nederlandse huisartsgeneeskunde? Welke vaardigheden dienen prioriteit te krijgen in het kader van deskundigheidsbevordering en toetsing?*
2. *Welke methoden en instrumenten zijn geschikt voor toetsing van technische vaardigheden van huisartsen en wat zijn hun meettechnische eigenschappen?*
3. *Wat is de effectiviteit van nascholing en toetsing van technische vaardigheden op het handelen in de praktijk? Heeft feedback over toetsingsresultaten effect op zelfbeoordeling?*
4. *Hoe is de acceptatie onder huisartsen van educatieve toetsing van technische vaardigheden? Welke vorm heeft de meeste voorkeur? Welke zijn de kosten en organisatorische randvoorwaarden voor toetsing van vaardigheden?*

In hoofdstuk 2 wordt het domein van medisch-technische vaardigheden voor de Nederlandse huisarts in kaart gebracht. Op basis van diverse huisartsgeneeskundige literatuurbronnen, werd een inventarisatie gemaakt, hetgeen uitmondde in een lijst met 263 vaardigheden, waarvan 85 % onder het Basistakenpakket gegroepeerd kan worden.
Uit de lijst met vaardigheden werden 80 vaardigheden geselecteerd die verondersteld werden prioriteit te hebben voor toetsing en nascholing. Deze lijst werd voorgelegd aan een twintigtal huisarts-coördinatoren deskundigheidsbevordering, die een rangordening wat betreft prioriteit aanbrachten. Met name vaardigheden op het terrein van oogheelkundige diagnostiek, onderzoek van het bewegingsapparaat, onderzoek bij hart- en vaatziekten, en vaardigheden in verband met thuiszorg, bleken hoge prioriteit te hebben. Deze vaardigheden vormden het uitgangspunt bij de keuze van onderwerpen in de diverse experimenten.

In hoofdstuk 3 wordt de literatuur besproken over vaardigheidstoetsing. Vaardigheidstoetsing is in de zeventiger jaren in het medisch curriculum geïntroduceerd in de vorm van het zogenaamde 'objective structured clinical examination' of examenvormen gebaseerd op 'standardized-patients', vanuit de behoefte om, naast kennistoetsing, praktisch medisch handelen te toetsen in situaties die de werkelijkheid zo goed mogelijk nabootsen, maar niet het nadeel hebben van de willekeur van het traditionele patient-gebonden examen. Geleidelijk heeft de examenvorm ook zijn weg gevonden naar specialisten-opleidingen. Het onderzoek waarmee de introductie van deze examenvorm gepaard ging heeft goede ondersteuning voor de validiteit opgeleverd bij een redelijke betrouwbaarheid. Er is in de literatuur echter ook kritiek dat de examenvorm te rigide en/of triviaal van inhoud zou zijn om competentie van meer ervaren artsen te toetsen.
In verband met de complexiteit van de organisatie bij vaardigheidstoetsing is er ook beperkt onderzoek verricht naar mogelijke alternatieven die goedkoper zijn en eenvoudig om toe te passen. Uit dit onderzoek bleek dat een kennistoets over vaardigheden een goede voorspeller is voor vaardigheidsbeheersing bij gevorderde studenten. Ook zelf-beoordeling is onderzocht op de voorspellende waarde voor vaardigheidsbeheersing. Dit leverde wisselende resultaten op, waarbij echter werd opgemerkt dat zelf-beoordeling een vaardigheid betreft die geleerd moet worden.
Op grond van deze bevindingen uit de literatuur werden behalve de vaardigheidstoets, ook kennistoetsing over vaardigheden en zelfbeoordeling in de uitwerking van de experimenten opgenomen, waarover in de verdere hoofdstukken wordt gerapporteerd.

In hoofdstuk 4 wordt verslag gedaan van een experiment waarbij een vaardigheidstoets (bestaande uit acht stations), een kennistoets over vaardigheden (125 vragen) en een zelfbeoordelingslijst (41 items) met elkaar werden vergeleken. Deelnemers aan het experiment waren 49 huisartsen en 47 huisartsen-in-opleiding. De gemiddelde scores tussen huisartsen en huisartsen-in-opleiding bleken niet significant te verschillen op de vaardigheidstoets en de

kennistoets over vaardigheden, terwijl de huisartsen wel significant hoger scoorden op de zelfbeoordelingslijst. De betrouwbaarheid van rangordening van deelnemers op basis van de vaardigheidstoetsscore was niet erg hoog (0.48) in vergelijking met de kennistoetsscore (0.68) en zelfbeoordelings-score (0.92). De samenhang tussen vaardigheidstoetsscore en kennistoets-score bleek hoog te zijn, terwijl de samenhang met de zelfbeoordeling een stuk lager was. Op basis van dit onderzoek werd geconcludeerd dat de kennistoets over vaardigheden een redelijk alternatief vormt voor de vaardigheidstoets, met name voor screeningsdoeleinden en onderzoek.

In hoofdstuk 5 wordt verslag gedaan van een tweede experiment, waarbij toetsing werd geïntegreerd in nascholing voor huisartsen. In dit experiment werd met name de construct-validiteit van de vaardigheidstoets nader onderzocht, alsmede de onderlinge samenhang tussen kennis en beheersing voor afzonderlijke vaardigheden. Daartoe werd een vaardigheidstraining (bestaande uit vier verschillende onderwerpen) gegeven aan 71 huisartsen. Het effect van de training werd gemeten met een vaardigheidstoets (vier stations) en een kennistoets over vaardigheden (60 vragen). De vaardigheidstoets liet een duidelijk trainingseffect zien voor alle onderwerpen, terwijl de kennistoets over vaardigheden dat voor drie onderwerpen liet zien. De samenhang tussen kennisscore en vaardigheidstoetsscore bleek echter laag te zijn voor al de afzonderlijke onderwerpen. Op grond hiervan werd geconcludeerd dat de vaardigheidstoets een goede construct-validiteit heeft, terwijl de samenhang tussen kennis over afzonderlijke vaardigheden en beheersing van die vaardigheden gering is.

In hoofdstuk 6 wordt het bij vaardigheidstoetsen veel gebruikte scoringssysteem - scorings-lijsten die worden ingevuld door observatoren - nader beschouwd. Dit gebeurt aan de hand van het station 'basale reanimatie', één van de vaardigheden uit het tweede experiment. De resultaten van het oordeel van een observator werden vergeleken met de resultaten van mechanische registratie verkregen tijdens het reanimeren. Beide scoringssystemen lieten een duidelijk trainingseffect zien bij de deelnemers. De observatoren bleken een goede onderlinge overeenstemming te hebben over het scoren van de diagnostische handelingen (0.77), terwijl de overeenstemming over de reanimatie handelingen beduidend lager was (0.56). Ook de correlatie tussen de scoringslijst en de mechanische registratie was laag (0.45) voor de reanimatie handelingen. Opvallend was ook dat de correlatie tussen de score voor diagnostische en die voor reanimatie handelingen laag was (0.22) voor beide scoringssystemen. Op basis van deze studie kan geconcludeerd worden dat de mechanische registratie een beter beeld geeft van de reanimatie handelingen dan de score door de observator. Omdat de score voor de reanimatie handelingen een slechte voorspelling geeft voor de diagnostische handelingen wordt echter een scoringssysteem aanbevolen dat zowel de diagnostische fase (middels een observator) als de reanimatie handelingen (middels mechanische registratie) omvat, zoals overigens ook veelal wordt gebruikt.

In hoofdstuk 7 wordt onderzocht in hoeverre een vaardigheidstraining, waarbij toetsing een integraal onderdeel van de cursus vormt, behalve een effect op de competentie ook een effect op het handelen in de praktijk heeft. De vaardigheidstraining omvatte een viertal vaardigheden: schouder injectie, cervixuitstrijk, fluor diagnostiek en fundoscopie bij diabetes mellitus. De deelnemende huisartsen (n=59) werden verdeeld in een interventie- en controlegroep. Alle deelnemers vulden een schriftelijke kennistoets over vaardigheden in en registreerden in hun praktijk gedurende twintig werkdagen hoe vaak zij de vier vaardigheden uitvoerden. De interventiegroep ontving vervolgens de vaardigheidscursus. De scores op de vaardigheidstoets na afloop van de training wezen op een redelijke tot goede vaardigheidsbeheersing. Daarop werd de schriftelijke toets herhaald voor alle deelnemers en volgde een tweede registratieperiode van twintig werkdagen.
De kennistoetsscores lieten een duidelijk trainingseffect zien van de vaardigheidscursus. Voor twee onderwerpen werd ook een trainingseffect op het handelen in de praktijk gevonden (schouder injectie en cervixuitstrijk), terwijl voor fluordiagnostiek en fundoscopie bij diabetes mellitus geen effect werd gevonden. Geconcludeerd wordt dat de effectiviteit van vaardigheidstraining om veranderingen in de praktijk te bewerkstelligen afhankelijk is van het onderwerp. Voor sommige vaardigheden is een training geschikt en voldoende om gewenste veranderingen te bewerkstelligen, terwijl voor andere vaardigheden waarschijnlijk meer complexe interventies nodig zijn.

In hoofdstuk 8 wordt nader ingegaan op zelfbeoordeling als methode om vaardigheidsbeheersing vast te stellen. Doel van het onderzoek waarover in dit hoofdstuk wordt gerapporteerd was na te gaan in hoeverre persoonlijke feedback over de scores op de kennistoets over vaardigheden en vaardigheidstoets de nauwkeurigheid van zelfbeoordeling verbeterde. De deelnemende huisartsen van het experiment, beschreven in hoofdstuk 7, werden aan het begin van de onderzoeksperiode, na drie maanden en na zes maanden getoetst op kennis (60 vragen) en vaardigheden (4 stations) en vulden een zelfbeoordelingslijst (22 items) in. Na verwerking van de resultaten ontvingen de deelnemers steeds persoonlijk feedback over hun scores. Na drie maanden bleken de kennistoets score en de zelfbeoordeling in de interventiegroep sterker gestegen in vergelijking met de controlegroep. Na zes maanden - nadat de controlegroep eveneens de training had gevolgd - waren de scores van beide groepen weer gelijk. De correlaties tussen de zelfbeoordeling en objectieve test scores waren vrij laag, zonder toename van de verklaarde variantie bij de achtereenvolgende meetmomenten.
Op basis van deze bevindingen lijken zelfbeoordelings scores in zekere mate in staat om op groepsnivo veranderingen in vaardigheidsbeheersing te meten. Echter, op individueel nivo, vormt zelfbeoordeling geen valide bron van informatie omtrent de beheersing van vaardigheden, en dit verbetert ook niet na feedback.

In hoofdstuk 9 wordt de waardering van de deelnemers voor vaardigheidstraining en -toetsing

beschreven, en wordt een overzicht gegeven van de kosten. Bij alle drie in eerdere hoofdstukken beschreven experimenten werden de deelnemers met een anonieme schriftelijke enquête gevraagd om hun mening over en waardering van de toetsing. De deelnemers waardeerden de toetsing als positief. De waardering was hoger bij huisartsen in vergelijking met huisartsen-in-opleiding, was hoger voor vaardigheidstoetsing in vergelijking met toetsing van kennis over vaardigheden, en was hoger bij experimenten waarbij de toetsing onderdeel uitmaakte van nascholing.
De kosten van vaardigheidstoetsing bedroegen ongeveer NLG 40,- per persoon per vaardigheid. Voor de combinatie van training en toetsing waren de kosten ongeveer NLG 100,-. De kosten zijn daarmee vergelijkbaar met de kosten van reguliere nascholing.

In hoofdstuk 10 worden de belangrijkste resultaten uit het onderzoek samengevat en van kanttekeningen voorzien. De volgende bevindingen worden als de meest belangrijke beschouwd. Ten eerste blijkt de vaardigheidstoets een goede validiteit te hebben voor het meten van technische vaardigheidsbeheersing bij huisartsen. Voor afzonderlijke vaardigheden kan een betrouwbare individuele beoordeling gegeven worden, maar voor het meten van het algemene vaardigheidsnivo zijn lange toetstijden nodig vanwege het inhoudsspecifieke karakter van vaardigheidsbeheersing. De kennistoets over vaardigheden vormt een goed alternatief voor een betrouwbare beoordeling van het individuele algemene vaardigheidsnivo, maar is niet geschikt voor het beoordelen van het vaardigheidsnivo voor afzonderlijke vaardigheden. Individuele zelfbeoordeling lijkt geen valide methode om vaardigheidsbeheersing vast te stellen.
Een tweede bevinding vormt het duidelijke effect van vaardigheidstraining op de competentie van de deelnemers, en de goede retentie die werd gevonden enkele maanden na de cursus. Bovendien werd bij twee van de vier vaardigheden waarbij dat was onderzocht, ook een effect van de nascholing op het handelen in de praktijk gevonden.
Een derde bevinding geldt de waardering en kosten van vaardigheidstoetsing in het kader van deskundigheidsbevordering voor huisartsen. De waardering blijkt hoog, en de meerkosten van toetsing zijn relatief bescheiden. De organisatie is relatief complex, maar goed realiseerbaar indien voldaan wordt aan een aantal randvoorwaarden. Daarom kan gesproken worden van een goede haalbaarheid van vaardigheidstoetsing als onderdeel van nascholing voor huisartsen.
Vervolgens worden een aantal methodologische kanttekeningen geplaatst bij de wijze waarop het onderzoek vorm heeft gekregen, en meer in het bijzonder bij verschillende aspecten van vaardigheidstoetsing.
Afgesloten wordt met aanbevelingen voor verder onderzoek, enerzijds gericht op een verdere onderbouwing van validiteit en educatieve waarde van vaardigheidstoetsing in nascholing voor huisartsen, anderzijds gericht op onderzoek naar een goede normstelling voor vaardigheidsbeheersing, met het oog op selectief gebruik van toetsing. De aanbeveling voor de praktijk betreft de noodzaak om goed doordachte ondersteuning te bieden wil men vaardigheidstoetsing in de praktijk ook ingang doen vinden.

Summary

Performance of technical clinical procedures constitute an important part of the work of general practitioners in the Netherlands. In this thesis technical clinical procedures of general practitioners are defined as patient related diagnostic and therapeutic procedures performed by a general practitioner. Examples of such procedures are physical examination, laboratory tests and minor surgery. In this thesis various methods for assessment of competence in technical clinical skills of general practitioners are explored in the context of continuing medical education.

In chapter 1 the background of the research project is described. Developments within the profession of general practitioners, as well as an increasing public demand for greater accountability with respect to quality of care, have provided the basis for the development of a quality system for Dutch general practice. The necessary elements for the quality system, such as national practice guidelines, assessment tools to ascertain how actual care compares with the guidelines, and methods for quality improvement for different dimensions of competence, are being developed. Technical clinical skills was considered one of the dimensions of interest.
Little information is available concerning the proficiency of practicing general practitioners with respect to the relevant technical procedures in general practice as agreed on by the profession. With respect to more recent technical developments in care which require training of new skills it is not known if and how general practitioners acquire these skills.
Related to the question which technical skills should be considered as essential to general practitioners, was the question how competence in technical clinical skills of general practitioners could best be measured, and how these methods could be integrated into the continuing medical education system. We hypothesized that inclusion of assessment and feedback to participants of continuing medical education would increase its efficacy. And finally the research project had to assess the feasibility of assessment, with regard to acceptability, cost and organisational requirements.

The research questions were:

1. *Which technical clinical skills are relevant for the general practitioner? Which skills should receive priority in continuing medical education and assessment?*
2. *Which methods are appropriate for assessment of technical clinical skills of general practitioners. What are the psychometric characteristics of these methods?*
3. *What is the effect of training and asessment of technical clinical skills on performance in practice? Does personal feedback enhance accuracy of self-assessment?*
4. *What is the acceptability among general practitioners of formative assessment of technical clinical skills? Which format has preference? What are the costs and organisational requirements for assessment of technical clinical skills?*

The subject of chapter 2 is which technical clinical skills are actually relevant for the general practitioner in the Netherlands. Based on a search in general practice literature a list of 263 technical skills was compiled. Of these 85% can be considered essential to the general practitioner.
From this list 80 skills were selected which were considered as having priority for assessment and continuing medical education. Subsequently twenty general practitioners involved as coordinators in continuing medical education rankordered these skills in terms of priority. Opthalmological diagnostic skills, physical examination of the locomotor system, diagnostic skills of the cardiovascular system, and skills related to intensive care at home, were categorized as having most priority. This priority list was used to select topics for the various experiments.

Chapter 3 reviews the literature on performance-based assessment. Performance-based assessment was introduced into the medical curriculum in the nineteen seventies as the so called 'objective structured clinical examination' or 'standardized patient-based' examination. This development was caused by the dissatisfaction with knowledge testing as dominant mode of competence assessment, and the subjectivity involved in most existing clinical assessments. Evaluation of performance of clinical tasks resembling real practice as close as possible, without the disadvantages of the traditional viva exam, was considered as a useful complement. The performance-based assessment format gradually also found it's way to postgraduate education. The research accompanying the introduction of this assessment method has provided support for good validity and reasonable reliability. However, the method has also been criticised for rigidity and/or trivialization of content, claiming it to be less suitable for assessment of competence of more experienced physicians.
Because performance-based assessment requires considerable resources some researchers have looked for alternative assessment methods which are less costly and easy to apply. A written knowledge test of skills showed good predictive validity for competence in technical clinical skills of advanced medical students. Self-assessment of technical clinical skills produced mixed results, with various authors commenting that self-assessment seemed to be a skill which has to be mastered. Based on the findings in the literature the performance-based test, the knowledge test of skills and self-assessment were included as assessment methods in the various studies reported in the following chapters.

Chapter 4 describes an experiment investigating the psychometric characteristics of three different methods for assessment of competence in technical clinical skills for general practitioners. A performance-based test (8 stations), a written knowledge test of skills (125 items) and a self-assessment questionnaire (41 items) on technical clinical skills were administered to 49 GPs and 47 trainees in general practice. The mean scores on the performance-based test and the written knowledge test of skills showed no substantial differences between general practiti-

oners and trainees, whereas the general practitioners scored higher on the self-assessment questionnaire. Norm-referenced reliability of the performance-based test was moderate (0.48) compared to the knowledge test of skills (0.68) and the self-assessment questionnaire (0.92). While the correlation of the score on the knowledge test of skills with the score on the performance-based test was moderately high, the score on the self-assessment questionnaire showed a rather low correlation with the performance-based test.
It was concluded that, although performance-based testing is obviously the best method to assess proficiency in hands-on skills, a written test can serve as a reasonable alternative, particularly for screening and research purposes.

Chapter 5 reports the results of a study with assessment integrated into continuing medical education for general practitioners. The study focussed on demonstrating construct-validity of the performance-based test for technical clinical skills, and exploring the correlation between performance and knowledge of specific skills. A one-day skills training was given to 71 general practitioners, covering four different technical clinical skills.
The effect of the training on performance was measured with a performance-based test (4 stations) using a randomized controlled trial design, while the effect on knowledge was measured with a written test (60 items) administered one month before and directly after the training. A training effect was found with the performance-based test for all four clinical skills. The written test also demonstrated a training effect for all but one skill. However, correlations between scores on the written test and on the performance-based test were low for all skills. It is concluded that construct validity of a performance-based test for technical clinical skills of general practitioners was demonstrated, while the knowledge test score showed to be a poor predictor of competence for specific technical skills.

In chapter 6 the scoring system which is predominantly used in performance-based assessment - with raters marking checklists while observing performance - is further explored. For cardiopulmonary resuscitation - one of the skills trained in the study reported in chapter 5 - checklist-based scores and mechanical recording scores were compared. Both checklist and recording strip based scores showed significant improvement after instruction, but only 37% were judged proficient according to the American Heart Association standards (checklist scoring), and 47% according to the recording print based scoring system, while raters judged 97% as satisfactory by general impression. Interrater reliability between observers was high for the diagnostic procedures (0.77) but much lower for CPR-performance (0.56). Comparison of checklist and recording print showed that the checklist was specific but not very sensitive in identifying poor performance for cardiac compression rate, since observers overestimated performance. The correlation for CPR-performance between checklist score and recording strip score was low (0.45), indicating that candidates were ranked differently. The correlation between diagnosis and performance score was low for checklist as well as recording print

(0.22), indicating that the score on diagnosis was a poor predictor for the score on performance of CPR. These results support the use of the recording manikin as compared with the use of a checklist for formative evaluation of basic life support skills. However, as proficiency in diagnosis and performance in CPR are poorly correlated, assessment of diagnosis using a checklist must be included. Therefore the combination of assessment by observers using a checklist for diagnostic procedures and the recording strip of the manikin for performance of CPR, as employed in most evaluation schemes, is recommended.

Chapter 7 reports the results of a study invesigating whether a short course of technical clinical skills with performance-based assessment integrated into the course has an effect on performance in practice. The course covered four different technical clinical skills (shoulder injection technique, PAP-smear, laboratory examination of fluor vaginalis, opthalmoscopic control in diabetes mellitus). Subjects were self-selected general practitioners (n=59), who were assigned to the intervention group (n=31) or control group (n=28) according to their preference for date of a course. The intervention group received the course three months after enrollment in the study, while the control-group received the training after the study period. Main outcome measures used were pre- and post-training scores on a knowledge test (60 items) and pre- and post-training performance of procedures in practice using a log-diary covering 20 days. Competence as measured with the knowledge test improved significantly as a result of the training, and skills test scores were satisfactory after training. A significant effect on performance was found for two out of four skills (shoulder injection and PAP-smear) whereas no effect could be demonstrated for the two other skills. It is concluded that a good degree of competence is a necessary but not always a sufficient condition for a physician to alter his performance in daily practice. While for some skills training seems adequate and sufficient to bring about desired changes, for other skills more complex interventions are needed.

Chapter 8 focusses on self-assessment as a method to determine competence in technical clinical skills. The purpose of the study reported was to ascertain if repeated personal feedback on knowledge test and performance-based test scores would enhance accuracy of self-assessment with regard to competence of technical clinical skills. Participants of the study described in chapter 7 completed a self-assessment questionnaire (22 items) covering four technical clinical skills and were assessed on relevant knowledge (60 items multiple choice test) at the start of the study period, again afther three months and after six months. A performance-based test (four stations) was administered after three and six months. After every assessment participants received personal feedback on their scores. At three months mean scores on the self-assessment questionnaire and knowledge test had increased significantly more in the intervention group compared to the control group, while after six months, after the control group had also received the training - no differences remained. Correlations between self-assessment rating and objective scores were low to moderate, with little overall improvement

of explained variance.
It is concluded that while self-assessment scores at the group level can be useful to some extent in measuring perceived changes in competence, individual self-assessment scores are an invalid source of information concerning competence of practicing physicians, and this does not improve significantly with regular feedback.

Chapter 9 reports on the acceptability of performance-based assessment and cost of the training and assessment applied in the various studies. Participants were asked to give their opinion on the assessment procedures and content of the course with an anonymous questionnaire. Participants valued the assessment as positive. Acceptability was higher among practicing physicians compared to trainees in general practice, higher for perfomance-based assessment compared to knowledge assessment, and higher when assessment was integrated into continuing medical education. The cost of performance-based assessment amounted to NLG 40,- per person per skill, and combined with training to NLG 100,- per person per skill. These costs are comparable to costs of regular CME in the Netherlands.

In chapter 10 the main findings are summarized and discussed. First, the performance-based test demonstrated good validity for assessment of technical skills of general practitioners. A reliable individual assessment is possible for specific skills. However for a reliable assessment of individual general competence in technical skills many stations are required and consequently long testing time is needed. This is due to the case-specificity of competence in technical clinical skills. The knowledge test of skills is considered a reasonable alternative for assessment of individual general competence in skills, whereas it is not suitable for individual assessment of seperate skills. Individual self-assessment does not provide valid results for competence in technical clinical skills. A second finding is the positive effect of skillstraining, including assessment, on competence with good retention various months after the course. Moreover, an effect of the course on performance in practice could be demonstrated for two out of four skills in which this was investigated.
A third finding concerns the acceptability and cost of assessment as part of continuing medical education. Acceptability is high and costs are reasonable. Organisation is complex, but feasible if certain requirements are met. Therefore feasibility of assessment of technical clinical skills in continuing medical education of general practitioners is favorable.
Subsequently some methodological problems are highlighted with respect to how the research was undertaken. This chapter is concluded with recommendations for research on validity and efficacy of performance-based testing in continuing medical education, and on developing standards for proficiency of technical clinical skills in general practice for selective assessment. Finally it is recommended to develop a sound strategy if implementation of performance-based assessment in CME is to be succesful.

Bijlagen

Overzicht Vaardighedenstations experimenten
Vaardighedenstations SVUH/WOK
Overzicht Kennis-over-vaardigheden vragen

Bijlage 1

Overzicht Vaardighedenstations experimenten

Experiment 1 (maart 1992)

- Fundoscopie
- Catheterisatie
- Mictieklachten
- Pijn op de borst
- Pijnlijke enkel
- Plaatsing IUD
- Reanimatie
- Verminderd gehoor

Experiment 2 (april 1993, maart 1994)

- Reanimatie
- Onderzoek schouder
- Infuus
- Injectie schouder

Experiment 3 (najaar 1994)

- Fundoscopie bij diabetes mellitus type II
- Injectie schouder
- Cervix uitstrijk
- Fluor/SOA-diagnostiek

Bijlage 2

Vaardighedenstations SVUH/WOK

A. Algemeen en niet gespecificeerd
Infuus inbrengen
Onderzoek van de pasgeborene
Recept schrijven A/B
Venapunctie

B. Bloed en bloedvormende organen
Laboratoriumonderzoek: hematologie
Laboratoriumonderzoek: BSE (Westergren)
Laboratoriumonderzoek: Hb (Spencer)

D. Tractus digestivus
Laboratoriumonderzoek: faecesonderzoek op Hemoglobine
Proctoscopie
Rectaal toucher

F. Oog
Onderzoek van het uitwendig oog en het voorste oogsegment
Fundoscopie
Oogboldrukmeting
Verwijderen corpus alienum
Visusonderzoek

H. Oor
Verwijdering cerumen gehoorgang
Verrichten van stemvorkproeven
Diagnostiek en behandeling van otitis externa
Beoordeling van het trommelvlies

K. Tractus circulatorius
Bloeddruk meten
Reanimatie zonder hulpmiddelen
Onderzoek decompensatio cordis
Onderzoek perifeer artierieel vaatlijden

L. Bewegingsapparaat
Onderzoek rug
Onderzoek schouder
Injectie schouder
Onderzoek knie
Onderzoek enkel

N. Zenuwstelsel
Onderzoek bij duizeligheid

R. Tractus respiratorius
Onderzoek bij dyspnoe
Onderzoek mond/keel
Indirecte laryngoscopie
Neusonderzoek
Astma: instructie inhalatiemedicatie

S. Huid en subcutis
Wondhechten
Excisie atheroomcyste
Behandelen van ulcus cruris
Laboratoriumonderzoek: diagnostiek nagel- en Huidmycosen

T. Endocr. kl./ metabolisme / voeding
Voetcontrole bij diabetes mellitus
Laboratoriumonderzoek: bepaling bloedglucose
Voedingssonde inbrengen

U. Urinewegen
Laboratoriumdiagnostiek: urine
Laboratoriumonderzoek: urinesediment
Laboratoriumonderzoek: de dipslide
Inbrengen eenmalige catheter man

W.Zwangersch/bevalling/anticonceptie
IUD inbrengen

X. Geslachtsorganen/borsten vrouw
Onderzoek borsten vrouw
Diagnostiek fluor vaginalis
Vaginaal en rectaal onderzoek
Laboratoriumonderzoek: gramkleuring
Maken van cervix uitstrijk

Y. Geslachtsorganen/borsten man
Laboratoriumonderzoek: de methyleenblauwkleuring
Rectaal toucher en proctoscopie

Bijlage 3

Overzicht Kennis-over-vaardigheden vragen

A	**Algemeen en niet gespecificeerd**	**18**
	Infuustoepassingen	10
	Bewusteloosheid	2
	Allergie	3
	Koorts kind	2
	Vaccinatie	1
B	**Bloed en bloedvormende organen**	**5**
	Bepaling BSE	3
	Bepaling Hb	2
D	**Tractus digestivus**	**6**
	afwijkingen mondholte/gebit	5
	Laboratoriumonderzoek	1
F	**Oog**	**42**
	Fundoscopie bij diabetes mellitus	15
	Fundoscopie algemeen	15
	Chalazion	2
	Onderzoek cornea/VOK	5
	Onderzoek visus	8
H	**Oor**	**10**
	Stemvorkproeven	2
	Inspectie trommelvlies	6
	Verwijderen cerumen	2
K	**Tractus circulatorius**	**36**
	Hartkloppingen	1
	Reanimatie	13
	Perifeer arterieel vaatlijden	15
	Problemen aderen	5
	Haemorrhoiden	2
L	**Bewegingsapparaat**	**56**
	Nekklachten	1
	Rugklachten	1
	Schouderklachten onderzoek	20
	Injectietechnieken schouder	10
	Pijnlijke elleboog/onderarm	5
	Pijnlijke pols/hand	1
	Heupklachten	2
	Knieklachten	8
	Fractuur onderbeen	1
	Pijnlijke enkel	7
N	**Zenuwstelsel**	**4**
	Krachtsverlies	2
	HNP-klachten	2
R	**Tractus Respiratorius**	**41**
	Indirecte laryngoscopie	10
	Dyspnoe	6
	Epistaxis	5
	CARA	20
S	**Huid en subcutis**	**26**
	Inspectie huidafwijkingen	4
	wondbehandeling	2
	abces	1
	Ulcus cruris	5
	Scabies	1
	Lyme-ziekte	2
	Haaruitval	1
	Schimmelinfectie huid	3
	Verbranding	3
	Mollusca/wratten	4
T	**Endocriene kl/metabolisme/voeding**	**10**
	Inbrengen voedingssonde	8
	Diabetes mellitus	2
U	**Urinewegen**	**10**
	Laboratoriumonderzoek	5
	Inbrengen catheter	5
W	**Zwangersch/bevalling/anticonceptie**	**10**
	IUD	5
	Onderzoek zwangere	1
	Begeleiding partus	4
X	**Geslachtsorganen/borsten vrouw**	**43**
	Onderzoek borsten vrouw	10
	Cervixuitstrijk	16
	Fluor vaginalis	11
	SOA	6
Y	**Geslachtsorganen/borsten man**	**11**
	SOA	4
	Afwijkingen penis/scrotum	3
	Fertiliteitsonderzoek	3
	Prostaatklachten	1

Curriculum vitae

Koos Jansen werd in 1956 geboren in Den Haag. Hij voltooide in 1975 zijn gymnasium B opleiding aan de scholengemeenschap 'De Breul' in Zeist. Hij studeerde daarna geneeskunde aan de Universiteit van Amsterdam waar hij in 1985 het artsexamen behaalde. Gedurende zijn studie was hij betrokken bij de oprichting van de Medicijnwinkel Amsterdam: een voorlichtingsproject over medicijnen voor consumenten. Vanaf 1979 was hij actief in de solidariteitsbeweging met de sandinistische revolutie in Nicaragua. Daarnaast was hij medeoprichter van de Stichting Werkgroep Medische Ontwikkelingssamenwerking (Wemos), die zich richt op de relatie tussen politieke en economische macht en gezondheid(szorg). Na het artsexamen werkte hij gedurende een jaar als arts-assistent in het Willem-Alexander ziekenhuis in Den Bosch. In 1987 volgde hij de éénjarige huisartsopleiding in Maastricht (opleider Felix Zwanikken in Heerlen). Vervolgens werkte hij via de organisatie Dienst over Grenzen twee jaar als tropenarts in Rama, Nicaragua. Teruggekeerd uit Nicaragua was hij gedurende vijf jaar als huisarts-onderzoeker verbonden aan de Werkgroep Onderzoek Kwaliteit Huisartsgeneeskunde ten behoeve van het onderzoeksproject dat in het proefschrift is beschreven. In die periode werkte hij ook als huisarts in het Gezondheidscentrum Hoensbroek in de praktijk van Geert-Jan van Schendel. Na afsluiting van het onderzoek verhuisde hij in 1996 naar Tilburg. Daar werkt hij sindsdien als huisarts. Verder is hij tijdelijk verbonden aan de Stichting Verenigde Universitaire Huisartsopleidingen in Utrecht in verband met een onderzoek naar vaardigheidsbeheersing van huisartsen-in-opleiding.